Planetary and Space Science

Editor-in-Chief: MILTON GREENBERG

Regional Editors:

D. R. BATES
A. I. LEBEDINSKY
S. K. MITRA
W. W. KELLOGG
D. F. MARTYN
M. NICOLET
F. L. WHIPPLE

Volume 7, 1961

PROCEEDINGS OF THE FOURTH AFBMD/STL SYMPOSIUM

ADVANCES IN

BALLISTIC MISSILE AND SPACE TECHNOLOGY

PERGAMON PRESS, INC.

NEW YORK · OXFORD · LONDON · PARIS

CONTENTS

PAGE

1. Instrumentation and Control

Design Factors and Performance Characteristics of a 500 VA Static Inverter 1
W. Steffe, *AC Spark Plug Division, General Motors Corporation*

A Practical Solution to the Arcing Problem at High Altitudes 11
Bradley R. Spink, *AC Spark Plug Division, General Motors Corporation*

Inertial Components to Supplement Astro-Fixes in Space Guidance 19
Charles J. Mundo, *Raytheon Manufacturing Company*

A Method of Damping Missile Structural Bending Modes for Autopilot Design 26
Lawrence E. Freed and Kenneth L. Miller, *Space Technology Laboratories, Inc.*

2. Space Guidance

A Navigation Theory for Round-Trip Reconnaissance Missions to Venus and Mars 40
Richard H. Battin and J. Halcombe Laning, Jr., *Instrumentation Laboratory, Massachusetts Institute of Technology*

A High Precision Stellar Navigator for Interplanetary Guidance 57
Charles D. Bock, *ARMA Division, American Bosch Arma Corporation*

A Guidance Technique for Interplanetary and Lunar Vehicles 64
William Schroeder and Clarence W. Pittman, *Space Technology Laboratories, Inc.*

Guidance and Control of Unmanned Soft Landings on the Moon 70
Sidney Darlington, *Bell Telephone Laboratories, Inc.*

A Fast Technique for Satellite Position Determination with Large Tracking Cameras 76
G. R. Miczaika and Nannielou H. Dieter, *Air Force Cambridge Research Center, Laurence G. Hanscom Field*

The Millstone Radar in Satellite and Missile Tracking 81
J. S. Arthur, J. C. Henry, G. H. Pettengill and P. B. Sebring, *Lincoln Laboratory, Massachusetts Institute of Technology*

A Simplified High Precision 200 Million Mile Tracking, Guidance and Communication System 94
James W. Crooks, Jr., *Convair (Astronautics) Division of General Dynamics Corporation*

A Vertical Azimuth Transfer System 108
Claude L. Emmerich and Robert E. Wisnieff, *Norden Laboratories, United Aircraft Corporation*

3. Trajectories

Generalized Trajectory Charts 118
A. D. Cohen, *Flight Propulsion Laboratory Department, General Electric Company*

PAGE

3. Trajectories—*continued*

A Method of Describing Miss Distances for Lunar and Interplanetary Trajectories 125
W. KIZNER, *Jet Propulsion Laboratory, California Institute of Technology*

Orbital Transfer for Satellites 132
SAMUEL P. ALTMAN, *The Martin Company*

Trajectories for Continuously Applied Propulsive Thrust 145
J. VICTOR HUGHES and GEORGE N. NOMICOS, *Republic Aviation Corporation*

Satellite Orbits Derived from a Gravitational Model of the Earth 154
PAUL H. SAVET, *ARMA Division, American Bosch Arma Corporation*

The Bi-Elliptical Transfer Between Co-Planar Circular Orbits 164
RUDOLPH F. HOELKER and ROBERT SILBER, *Army Ballistic Missile Agency, Aeroballistics Laboratory*

Applications of Dynamic Programming to Space Guidance, Satellites, and Trajectories 176
RICHARD BELLMAN, STUART DREYFUS and ROBERT KALABA, *The RAND Corporation*

4. Computers

Missile Simulation on an Electronic Digital Computer 179
A. HASKELL EPSTEIN and NORMAN W. PETERSON, *Space Technology Laboratories, Inc.*

BIAX High Speed Magnetic Computer Element 184
C. L. WANLASS and S. DEAN WANLASS, *Aeronutronic, a Division of Ford Motor Company*

5. Reliability

The Air Force Approach to Reliability in Ballistic Missile Weapon Systems 208
VINCENT J. BRACHA, *AFBMD*

Reliability Evaluation of Large Solid Rocket Engines During Engineering Development 217
M. LIPOW and D. K. LLOYD, *Space Technology Laboratories, Inc.*

Techniques in the Demonstration of Reliability Attainment 230
EDWIN D. KARMIOL and JOHN S. YOUTCHEFF, *Missile and Space Vehicle Department, General Electric Company*

Analysis of Random Failures 242
G. P. ANDERSON and L. E. PETERSON, *Remington Rand Univac, Division of Sperry Rand Corporation*

6. Space Communications

Radio Communication by Means of Satellites 254
C. C. CUTLER, *Bell Telephone Laboratories, Inc.*

A Study of UHF Space Communications Through an Aurora Using the Moon as a Reflector 272
R. P. INGALLS, J. C. JAMES and M. L. STONE, *Lincoln Laboratory, Massachusetts Institute of Technology*

Analysis and Prediction of Radio Signal Interference Effects Due to Ionized Layer Around a Re-Entry Vehicle 286
WILLIAM C. TAYLOR, *Lockheed Missiles and Space Division*

PAGE

7. Man in Space

The Dilemma of Man in Space 301
J. S. BRADY, *Space Technology Laboratories, Inc.*

A Hypothetical Mission to Space in a Three-Man Sealed Cabin 309
J. I. R. BOWRING and B. P. EBERT, *Wright-Patterson AFB, Ohio*

Space Simulators—Prelude to Manned Space Flight 324
A. B. THOMPSON and W. B. LUTON, *Vought Astronautics Division of Chance Vought Corporation*

Space Flight Acceleration Simulator 335
CHARLES L. BARKER, JR., *Army Ballistic Missile Agency, Army Ordnance Missile Command*

Closed Ecology Problems in a Space Cabin Simulator 345
JEAN R. NELSON, *Minneapolis-Honeywell Regulator Company, Aeronautical Division*

Package Experiments for Bioastronautics Research 356
LAUREL VAN DER WAL, *Space Technology Laboratories, Inc.*

A Partially Closed Cycle Life Support System for Long-term Space Flight 370
J. J. KONIKOFF, *Missile and Space Vehicle Department, General Electric Company*

Selection of Physiologic Measurements 378
FRANCIS BERCHMANS QUINN, JR., and VERNE O. BRECHNER, *Department of Surgery, University of California, Medical Center, Los Angeles, California*

Impact Tolerance of Mammals 386
M. KORNHAUSER and R. W. LAWTON, *Missile and Space Vehicle Department, General Electric Company*

Human Engineering or Engineering of the Human—Which? 395
ROBERT S. POGRUND, *Space Technology Laboratories, Inc.*

8. Space Experiments

Composition of Radiation Trapped in the Geomagnetic Field at Altitudes up to 1000 Kilometers 411
FRANCIS E. HOLLY, *Air Force Special Weapons Center*, and RICHARD G. JOHNSON, *Lockheed Missiles and Space Research Laboratory*

Measurements of Flux of Small Extra-Terrestrial Particles 417
HERBERT A. COHEN, *Air Force Cambridge Research Center, Laurence G. Hanscom Field*

Deduction of Vertical Thermal Structure of a Planetary Atmosphere from a Satellite 423
JEAN I. F. KING, *Missile and Space Vehicle Department, General Electric Company*

Thermal Impact—The Mechanical Response of Solids to Extreme Electromagnetic Radiation 427
J. E. MICHAELS, *Missile and Space Vehicle Department, General Electric Company*

A Dynamical Model for Investigation of the Downward Transfer of Energy Introduced by Upper Atmospheric Heating 434
LOUIS BERKOFSKY and RALPH SHAPIRO, *Air Force Cambridge Research Center, Laurence G. Hanscom Field*

Recent Information Gained from Satellite Orientation Measurement 445
ROBERT J. NAUMANN, *Army Ballistic Missile Agency, Research Projects Laboratory*

Review of the ARDC Model Atmosphere, 1959 454
K. S. W. CHAMPION, R. A. MINZNER and H. L. POND, *Air Force Cambridge Research Center, Laurence G. Hanscom Field*

Planetary and Space Science

CUMULATIVE INDEXES

Volumes 6–8
1961

Editor-in-Chief: MILTON GREENBERG

Regional Editors:

D. R. BATES
W. W. KELLOGG
D. F. MARTYN
T. NAGATA
W. ORR ROBERTS
H. FRIEDMAN
A. I. LEBEDINSKY
S. K. MITRA
M. NICOLET
F. L. WHIPPLE

PERGAMON PRESS

NEW YORK · OXFORD · LONDON · PARIS

Subject Index

Absorption
- microwave radiation in a plasma, **6**, 105.
- radiowaves in ionosphere: **8**, 151.
- soft X-rays by aluminium: **8**, 169

Air, high temperature
- electron concentration in closed form: **6**, 79.
- collision frequency: **6**, 24.
- cross-sections for constituents: **6**, 24

Air, seeded: **6**, 79, 172.
Airglow, artificial, measurement of diffusion coefficient: **8**, 35.
Antennae
- ambipolar diffusion, effect on breakdown: **6**, 149
- power handling at high altitudes: **6**, 142
- radiation from slotted-sphere: **6**, 116
- radomes, microwave breakdown at high altitudes: **6**, 155.
- R.F. breakdown in presence of sheath: **6**, 149.

Arcing, at high altitudes: **7**, 11
Atmosphere of Earth
- abundance of N_2, **8**, 9.
- ARDC 1959 model, review: **7**, 454.
- model for downward transfer of energy: **7**, 434.
- thermal structure of ozonosphere from satellite measurements: **7**, 423.

Aurora
- circular symmetry in geomagnetic equatorial plane: **8**, 142.
- excitation of high altitude red arcs: **8**, 59.
- excitation of subvisual OI arcs at middle latitudes: **8**, 198.
- observations during I.G.Y. and I.G.C.: **8**, 125.
- observations in Alaska including high altitude red arcs: **8**, 49.
- precipitation of trapped particles: **8**, 13.
- transmission of UHF signals through: **7**, 272.

Book reviews
- *An introduction to celestial mechanics* by T. E. Sterne: **8**, 122.
- *Cosmic radio waves* by I. S. Shklovskii, trans. by R. B. Rodman and C. M. Varsavsky: **8**, 122.
- *Physics of the solar chromosphere* by Richard N. Thomas and R. Grant Athay: **8**, 122.

Carbon dioxide, in atmosphere of Venus: **8**, 23.
Collision frequencies, high temperature air: **6**, 24.
Communications, by satellites: **7**, 254.
Computers
- high speed memory element: **6**, 184.
- random failures in components: **7**, 242
- simulation of missiles: **7**, 179.

Cosmic rays, measurement on geophysical rockets: **8**, 165.

Diffusion
- ambipolar, effect on antenna breakdown: **6**, 149.
- coefficients, determined from sodium vapour trails: **8**, 35.

Electron concentration in high temperature air: **6**, 79.
Exosphere, model of ion distributions in: **8**, 71.

Guidance
application of dynamic programming: **7,** 176.
control of soft landings on the Moon: **7,** 70.
high precision stellar navigator: **7,** 57
homing: **7,** 125
inertial components to supplement astro-fixes: **7,** 19.
interplanetary and lunar vehicles: **7,** 64, 94.
miss distances for lunar and interplanetary trajectories: **7,** 125.
orbital transfer for satellites: **7,** 132.
orbits of space rockets in the direction of the Moon: **8,** 107.
problems in interplanetary space: **8,** 97.
reconnaissance missions to Venus and Mars: **7,** 40.
static inverter: **7,** 1.
trajectories for continuous propulsive thrust: **7,** 145.
vertical azimuth transfer system: **7,** 108.

Hydroxyl emission, nightglow: **8,** 197.
Hypersonics
collision frequencies in high temperature air: **6,** 24.
effect of errors in rate constants on calculated shock layer electron densities: **6,** 74.
electron concentration in closed form for high temperature air: **6,** 79.
ionized shock layer for blunt shapes: **6,** 10.
plasma data of equilibrium air: **6,** 47.
radio signal interference due to ionized layer: **6,** 1; **7,** 286.
rate of ionization behind shock waves: **6,** 94.

Ionization
non-equilibrium, in shock waves: **6,** 74.
rate behind shock waves in air: **6,** 94.
rate in lower atmosphere: **8,** 9.
Ionosphere
D-region, origin of: **8,** 200.
electron concentration: **8,** 183.
positive ion concentration from Sputnik III: **8,** 179
Ions, in exosphere: **8,** 71.

Kepler problem, equations of disturbed motion: **8,** 79.

Micrometeorites
flux measurement: **7,** 417.
on Sputnik III: **8,** 82.
Micro-organisms, viability in ultra-high vacuum: **8,** 30.
Microwaves
radiation, absorption in plasma: **6,** 105.
techniques for hypersonic ballistic ranges: **6,** 186.
Missiles
damping of structural bending: **7,** 26.
performance of static inverter: **7,** 1.
radar tracking: **7,** 81.
reliability: **7,** 208, 217, 230.
simulation by a digital computer: **7,** 179.
trajectory charts: **7,** 118.
Moon
expansion and contraction: **8,** 1.

linear array of small craters: **8**, 121
orbits of rockets: **8**, 107.
soft landings on: **7**, 70.

Nightglow, hydroxyl emission: **8**, 197.
Nitrogen, molecular, abundance in terrestrial atmosphere: **8**, 9.

Plasma
coated sphere, echo area: **6**, 133.
data of equilibrium air: **6**, 47.
from seeded cyanogen flame, propagation in: **6**, 172.
non-uniform, interaction with microwave radiation: **6**, 105.
propagation properties: **6**, 1.
sheath, on spherical satellite, radar cross-section: **6**, 123.
sheath, physical properties around projectiles: **6**, 196.
sheath, propagation through, in hypersonic low density wind tunnel: **6**, 180.
waves, attenuation of: **6**, 1.
waves, in a magnetic field: **6**, 100.
WKB approximation in: **6**, 105.

Radiation, solar, rocket measurements: **8**, 120.
Reconnaissance, mission to Venus and Mars: **7**, 40.
Re-entry
effects of impact: **7**, 386.
radiation from slotted-sphere antenna with plasma sheath: **6**, 116.
radio signal interference due to ionized layer: **6**, 1; **7**, 286.
telecommunications during: **6**, 207.
R.F. fields, random, measurement using short electric and magnetic dipoles: **8**, 68.

Satellites
bi-elliptical transfer between coplanar circular orbits: **7**, 164.
communications by means of: **7**, 254.
determination of Earth's gravitational potential: **8**, 43.
effect of Earth's oblateness on rotational motion: **8**, 173.
Hohmann transfer: **7**, 132, 164.
illumination conditions and periods in shadow and sunlight: **8**, 159.
impact of meteoric particles on Sputnik III: **8**, 82.
orbital transfer: **7**, 132.
orbits derived from gravitational model of the Earth: **7**, 154.
orientation effects: **7**, 445.
path in normal gravitational field of the Earth: **8**, 86.
positive ion concentrations from Sputnik III: **8**, 173.
radar cross-section: **6**, 123.
Shock layer
blunt shapes: **6**, 10.
non-equilibrium ionization in: **6**, 74.
transmission of e.m. radiation through **6**, 1.
Sodium, vapour trails: **8**, 35.
Solar, radiation, rocket measurements: **8**, 120.
Space flight
bioastronautics research: **7**, 356.
closed cycle life support system: **7**, 370.
ecology problems: **7**, 345.
effects of impact: **7**, 386.
hypothetical orbital flight: **7**, 309.

man in space: **7**, 301.
micro-organisms: **8**, 30.
physiological effects: **7**, 378, 395.
space simulators: **7**, 324, 335.
Thermal impact: **7**, 427.
Tracking
generalized trajectory charts: **7**, 118.
high precision 200 million mile system: **7**, 94.
photographic satellite position determination: **7**, 76.
radar, for satellites and missiles: **7**, 81.

Van Allen radiation
composition of lower belt: **7**, 411.
longitudinal dependence: **8**, 13.
Velocity, corrections for space vehicle: **7**, 40.
Venus, CO_2 content of atmosphere: **8**, 23.

Waves, plasma, in magnetic field: **6**, 100.
Whistlers, dispersions for an ionospheric model: **8**, 71.
W.K.B. approximation, in a plasma: **6**, 105.

Author Index

(B signifies book review and R research note)

Allis, W. P. **6**, 100
Altman, S. P. **7**, 132
Anderson, G. P. **7**, 242
Arthur, J. S. **7**, 81

Bachynski, M. P. **6**, 24
Baldwin, K. M. **6**, 207
Barker, C. L., Jnr. **7**, 335
Basset, O. E. **6**, 207
Battin, R. H. **7**, 40
Bellman, R. **7**, 176
Berkofsky, L. **7**, 434
Bock, C. D. **7**, 57
Bowring, J. I. R. **7**, 309
Bracha, V. J. **7**, 208
Brady, J. S. **7**, 301
Brechner, V. O. **7**, 378
Brown, R. Hanbury, **8**, 123 (B)
Brueckner, K. A. **6**, 105
Brueschke, E. E. **8**, 30

Champion, K. S. W. **7**, 454
Chuan, R. L. **6**, 180
Cohen, A. D. **7**, 118
Cohen, H. A. **7**, 417
Cook, G. E. **8**, 122 (B)
Crooks, J. W. **7**, 94
Cutler, C. C. **7**, 254

Darlington, S. **7**, 70
Deehr, C. S. **8**, 49
Dieter, N. H. **7**, 76
Dolph, C. L. **6**, 123
Dreyfus, S. **7**, 176

Ebert, B. P. **7**, 309
Emmerich, C. L. **7**, 108
Epstein, A. H. **7**, 179

Fielder, G. **8**, 1
Fiskin, J. M. **6**, 47
Freed, L. E. **7**, 26

Ghosh, S. N. **8**, 9
Green, R. B. **6**, 133
Greyber, H. D. **6**, 105
Gringauz, K. I. **8**, 183

Hawthorne, E. I. **6**, 207
Henry, J. C. **7**, 81
Hibbs, A. R. **8**, 121 (R)
Hochstim, A. R. **6**, 79
Hoelker, R. F. **7**, 164
Holly, F. E. **7**, 411
Huber, P. W. **6**, 172
Hughes, J. V. **7**, 145
Hugill, J. **8**, 68
Hultqvist, B. **8**, 142

Ingalls, R. P. **7**, 272
Inn, E. C. Y. **8**, 201 (R)
Istomin, V. G. **8**, 179

James, J. C. **7**, 272
Johnson, R. G. **7**, 411
Johnston, T. W. **6**, 24

Kalaba, R. **7**, 176
Kaplan, L. D. **8**, 23
Karmiol, E. D. **7**, 230
Keenan, P. P. **6**, 155
King, J. I. F. **6**, 105; **7**, 423
Kislik, K. D. **8**, 86
Kizner, W. **7**, 125
Klein, M. M. **6**, 105
Konikoff, J. J. **7**, 370
Kornhauser, M. **7**, 386
Krasovskii, V. I. **8**, 125, 197 (R)

Langberg, E. **6**, 207
Laning, J. H. **7**, 40
Lawton, R. W. **7**, 386
Lin, S. C. **6**, 94
Lipow, M. **7**, 217
Lloyd, D. K. **7**, 217
Loughnan, C. J. **8**, 13
Lur'e, A. I. **8**, 79
Luton, W. B. **7**, 324

Macdonald, A. D. **6**, 149
McNally, D. **8**, 122 (B)
Maiden, C. J. **6**, 196
Marini, J. W. **6**, 116
Michaels, J. E. **7**, 427
Miczaika, G. R. **7**, 76
Miller, K. L. **7**, 26
Minzner, R. A. **7**, 454
Morita, T. **6**, 142
Mundo, C. J. **7**, 19

Nauman, R. J. **7**, 445
Nazarova, T. N. **8**, 82
Nelson, J. R. **7**, 345
Nomicos, G. N. **7**, 145

Papa, R. J. **6**, 100
Patterson, T. N. L. **8**, 71
Peters, L. **6**, 133
Peterson, L. E. **7**, 242
Peterson, N. W. **7**, 179
Pettengill, G. H. **7**, 81
Pittman, C. W. **7**, 64
Pogrund, R. S. **7**, 395
Pond, H. L. **7**, 454
Pounds, K. A. **8**, 120 (R)
Primich, R. I. **6**, 186
Quinn, F. B. **7**, 378

Raushenbakh, B. V. **8**, 97
Rees, J. A. **8**, 35
Rees, M. H. **8**, 49, 59, 198 (R)
Ridyard, H. W. **6**, 10
Rudakov, V. A. **8**, 183

St. Pierre, C. **6**, 196
Sarychev, V. A. **8**, 173
Savet, P. H. **7**, 154
Scharfman, W. E. **6**, 142
Schroeder, W. **7**, 64
Sebring, P. B. **7**, 81
Sedov, L. I. **8**, 107
Shafer Yu. G. **8**, 165
Shapiro, R. **7**, 434
Sharma, K. D. **8**, 9
Shkarovskii, I. P. **6**, 24
Silber, R. **7**, 164
Sisco, W. B. **6**, 47
Smith, D. E. **8**, 43
Spink, B. R. **7**, 11
Steffe, W. **7**, 1
Stone, M. L. **7**, 272
Suess, R. H. **8**, 30

Taylor, W. C. **6**, 1; **7**, 286
Thompson, A. B. **7**, 324
Tokar, E. N. **8**, 97

Wal, L. Van Der **7**, 356
Wanlass, C. L. **7**, 184
Wanlass, S. D. **7**, 184
Weil, H. **6**, 123
Whitehead, J. D. **8**, 151
Whitmer, R. F. **6**, 149
Willard, M. **8**, 30
Wisnieff, R. E. **7**, 108

Yarygin, A. V. **8**, 165
Yatsunskii, I. M. **8**, 159
Youtcheff, J. S. **7**, 230

DESIGN FACTORS AND PERFORMANCE CHARACTERISTICS OF A 500 VA STATIC INVERTER*

W. STEFFE
AC Spark Plug Division, General Motors Corp.

Abstract—Design and performance characteristics of all silicon static inverters are presented. One unit delivers in excess of 25 VA/lb and will operate with heat sink temperatures up to 100°C. Output voltage regulation, better than 0·5 per cent, is obtainable by means of a strong feedback loop which includes a highly temperature compensated reference amplifier. Special switching power amplifiers produce reasonable efficiency and low distortion. Appropriately packaged the unit is well suited for missile operation.

INTRODUCTION

With ever increasing demands made on missile guidance system capabilities, the performance required of missile guidance systems power sources, such as inverters, also increases. The event of higher power silicon control devices offers opportunities to design static inverters with performance superior to rotary inverters. The use of silicon devices also permits operation at temperatures exceeding the limit imposed by germanium. The design of a typical all-silicon static inverter was undertaken in order to meet a particular requirement for a two phase static inverter, and to demonstrate the degree of accuracy obtainable.

Basically, there are two different silicon devices available that are capable of the desired performance; the silicon controlled rectifier (SCR) and the silicon power transistor. Both types have their advantages and disadvantages and will be discussed in more detail. The silicon transistor approach was chosen because of its low distortion and less critical operating requirements. The relatively high cost of the inverter, reflecting the present high price of silicon power transistors, can be justified only in view of the performance obtained.

ORIGIN OF THE PROBLEM

During the past years rotary inverter performance has been improved and 1 per cent voltage regulation is commonly obtained. Weight has decreased so that continuous ratings of 30 VA/lb are available in equipment with a 500 VA capacity. Light weight has been obtained only with the sacrifice of efficiency. Efficiency lower than 50 per cent is quite common.

Static inverters were originally designed using germanium transistors. Efficiencies of 85 per cent are not uncommon in units comparable to rotary machines in weight. Regulation to 1 per cent is frequently obtained. Unfortunately, germanium imposes a maximum heat sink temperature around 70°C. With limited quantities of relatively warm air available for heat sink cooling this condition makes large heat sinks mandatory. For the reliability required, and space allowed, the use of germanium power transistors is not particularly desirable in some missile environments. The frequently used switching designs also tend to become more bulky and less efficient as requirements are made more stringent by requiring lower harmonic distortion content in the output.

Recently, higher power silicon devices have become available on the commercial market. The silicon controlled rectifiers with ratings up to 50 A and power transistors of 5 A are presently available. Transistors with 20 A capabilities are expected late this year. These devices can operate with heat sink temperatures of 100°C and still handle appreciable power. This paper concerns itself with an investigation of the feasibility of utilizing these silicon devices for static inverter use. It is intended to outline the

* Work accomplished on AF 04(645)-19 for U.S. Air Force.

application of these devices and demonstrate the high degree of precision output control obtainable by the use of silicon power transistors.

The precision performance required of the static inverter described in this report requires a design approach different from the approaches feasible for generation of power with less stringent regulation and stability requirements. Efficiency and weight–power ratio are therefore not comparable to high power inverters primarily aimed toward direct replacement of rotary inverters. The approach outlined in this report is applicable only where precise power characteristics are required and is uniquely suited for the generation of accurate alternating current power up to 1 kVA from aircraft type direct current power sources.

APPROACH

The design requirements for a missile guidance system inverter were established.

Highlights of these specifications were:

(1) Power output of the unit to be 500 VA total, 175 VA in one phase and 325 VA in the second phase.

(2) R.M.S. voltage regulation better than plus or minus 0·3 per cent for the following conditions of load:

325 VA maximum, phase *A*. 195 VA minimum, phase *A*. 170 VA maximum, phase *B*. 105 VA minimum, phase *B*. Load unbalance less than 10 per cent.

(3) Heat sink temperature range 0°–100°C. Components not mounted on heat sink: 0°–85°C.

(4) Efficiency better than present rotary inverter of 45 per cent; design goal 75 per cent.

(5) Distortion less than 3 per cent total harmonic.

(6) Input power 25–29 V d.c.

(7) The inverter must withstand short circuit and no load conditions for up to 10 sec without adverse effects upon the life of the unit.

(8) Phase displacement between the two phases to be $90° \pm 1°$.

The main work effort was directed toward meeting the requirements outlined above, but an effort was made to establish the best characteristics obtainable. In order to obtain the regulation required the output impedance of the output stages must be lowered by means of an additional voltage control loop. The problem which arises here is to properly sense the magnitude of the output voltage. Due to the wide ambient temperature range specified the detector must be inherently temperature insensitive if the use of an oven is to be avoided. True r.m.s. detectors such as thermocouples or thermistors do not meet these requirements because they all are necessarily temperature sensitive. Approximate r.m.s. detectors using diodes also tend to be excessively temperature sensitive and complex. Average detectors can be well temperature compensated and are simple in circuitry. Unfortunately an average detector can be substituted for a r.m.s. detector only if the distortion components remain relatively low. It becomes necessary, therefore, to use one of two possible approaches: (i) use of a true r.m.s. detector permitting the use of power output stages meeting the 3 per cent distortion requirement; (ii) use of an average detector and inverter power output stages with distortion sufficiently low and constant to avoid excessive detector errors. The second approach was chosen. A closer look at the detector and the voltage control loop is therefore required to arrive at the maximum distortion permissible in the power output stages.

CONTROL CIRCUIT

The block diagram of Fig. 1 serves to illustrate the control circuit operation. The loop gain is kept very high, approximately 60 dB, to obtain a very low effective output impedance from the inverter. The high degree of feedback causes the output voltage to be almost entirely dependent upon the stability of the reference element and the detector circuits. Both the detector circuit and the reference amplifier

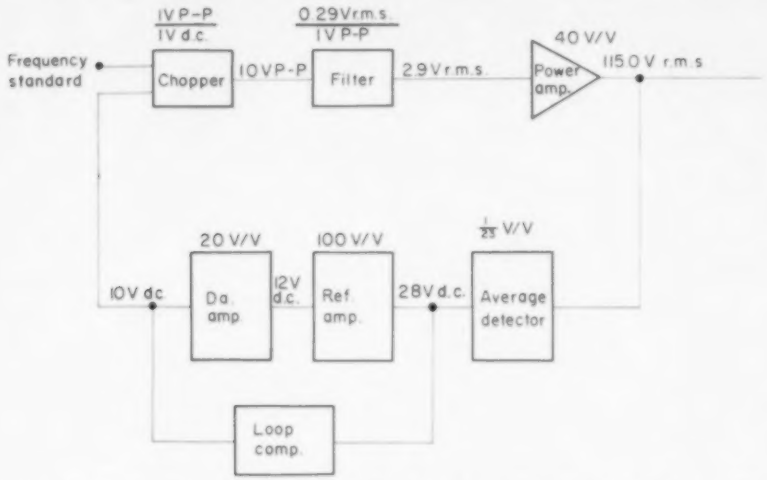


Fig. 1. Voltage control loop for one phase of static inverter. Loop gain is approximately 60 dB. Indicated transfer functions are nominal small signal values.

circuit have been designed with great care. This was done particularly in respect to temperature stability and stability in respect to time.

Detector

Due to the location of the detector circuit in the feedback network, any error arising in it is directly reflected to the output voltage. The detector circuit has to function precisely to assure the best possible voltage regulation for the inverter. The detector must meet the following requirements:

(1) It must convert an r.m.s. a.c. signal to an adjustable d.c. level directly proportional to the a.c. signal.

(2) The output voltage must contain less than 50 mV of ripple to avoid overloading the subsequent circuits.

(3) The phase and frequency response must be compatible with loop requirements.

(4) The transfer function must be temperature and time stable.

(5) The effects of distortion in the output waveform must be predictable.

The detector circuit of Fig. 2 meets these

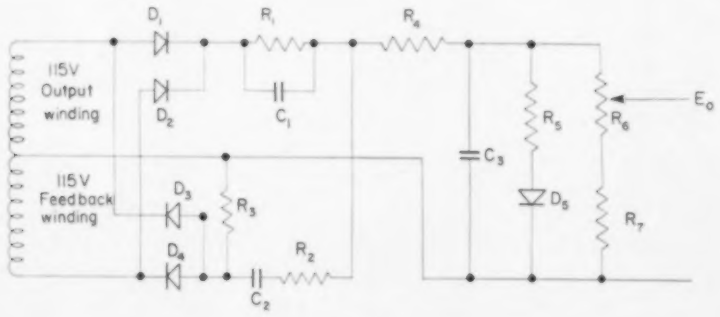


Fig. 2. Average detector with ripple cancelling circuit and temperature compensation.

$$f_0 = \frac{1}{2\pi RC} \qquad C_2 = 2C_1 \qquad R_2 = 0{\cdot}5\,R_1$$

requirements to a remarkable degree. A full wave rectifier produces positive and negative output voltages which are applied to a noise cancelling circuit ($R1$–$C1$, $R2$–$C2$ and $R3$ which serves as load on the otherwise unloaded negative rectifier). $R4$ and $C3$ combine to form an RC filter with a time constant sufficiently small to avoid phase shift within the response of the control loop. Ripple voltages well below 1 per cent have been obtained with a circuit pass band (45° point) of 50 cps. Temperature stability has been obtained by cancelling the temperature coefficients of the rectifying diodes $D1$, $D2$, $D3$ and $D4$ with the coefficient of $D5$. The coefficient of these diodes is approximately 2 mV/°C or 170 mV over the ambient range expected at the diode location. This represents an error of approximately 0·16 per cent if not compensated. With the use of temperature compensating diode $D5$, this error has been reduced to less than 0·02 per cent. Temperature and time stability are further assured by the use of high grade teflon capacitors and wire-wound resistors. The maximum effect of distortion components in the inverter output can be analyzed rather easily for effects on an average detector. In order to demonstrate the maximum effect of harmonics upon the detector output an example will be used. To a pure sine wave of magnitude 100 V distortion will be added, 1 per cent second harmonic and 1 per cent third harmonic. While the change in the r.m.s. voltage is very small, given by the square root of the sum of the squares, the average voltage is effected differently. During one half cycle of the fundamental the second harmonic goes through a full cycle, that is both a negative and a positive half cycle, with an average value of zero. A second harmonic therefore has no effect upon the detector output. The third harmonic swings through three half cycles during the half cycle of the fundamental, and thereby produces two half cycles of opposing polarity with no net average value, plus one half cycle which can produce a change in the average value. The magnitude of the contribution is dependent upon the relative phase position of the harmonic. The maximum change in average value produced by 1 per cent third harmonic is one third of 1 per cent. This reasoning can be extended to show that even harmonics have no effect upon the average value, while odd harmonics have maximum effects reduced in magnitude by the order of the harmonic. To arrive at a figure for the maximum error to be expected from the average detector the theory outlined above was applied to the distortion spectrum typical for a high power amplifier. The results are shown in the table in Fig. 3. A constant distortion will cause a calibration error only, and may therefore be adjusted out. If the distortion varies, due to loading, environment or supply conditions, a 1 per cent total harmonic distortion change can produce up to one third of 1 per cent change in the output of the average detector, and therefore, in the inverter a.c. output. It is assumed here that the

Order of harmonic	Per cent of distortion	Maximum effect on av. value	Net per cent contribution to av. value	Net per cent contribution to r.m.s. value
2	0·54	0	0	0·0015
3	0·79	1/3	0·263	0·0031
4	0·31	0	0	—
5	0·45	1/5	0·090	—
6	0·12	0	0	—
7	0·15	1/7	0·021	—
	r.m.s. Total 1·12		0·374	r.m.s. Total 0·00626

Fig. 3. Typical distortion spectrum of a power amplifier and the effects upon r.m.s. and average values.

distortion distribution remains approximately as shown in Fig. 3 and that the relative phase position of the harmonics is the worst possible. A factor of this magnitude can certainly not be neglected if tight regulation of the r.m.s. value is required. Total harmonic distortion of 1 per cent appears to be the desired maximum limit to meet the voltage regulation requirement of plus or minus 0·3 per cent. With 1 per cent maximum distortion, the variable distortion components will be less than 0·5 per cent as indicated by test results. This will produce a maximum regulation error of 0·15 per cent from this source. If properly analyzed, a corrective signal may be derived from the distortion components. This approach was not chosen here because of its additional complexity.

As shown on the block diagram (Fig. 1), the detector output is fed into a high gain compensated d.c. amplifier, with a reference amplifier as the first stage. The reference amplifier must produce an output signal directly proportional to a deviation from the nominal detector output voltage. To accomplish this a selected type 3*N*44 reference amplifier is used. The guaranteed temperature coefficient of this commercially available unit is 0·002 per cent/°C. The unit consists of a transistor and a reference element mounted in a common heat sink. A constant collector current requires a voltage from the base to the low side of the reference element varying by not more than 20 p.p.m./°C. This temperature coefficient of 0·002 per cent/°C results in approximately 0·17 per cent error for the full ambient range. Proper packaging can reduce this figure. For very tight voltage control some form of temperature control is required. The d.c. gain of the reference amplifier is approximately 100 V/V. The reference amplifier is followed by a temperature compensated two-stage d.c. amplifier with high input impedance and a gain of 20 V/V. A feedback network around the reference amplifier and the d.c. amplifier introduces the response curve shaping required for stability of the voltage control loop. The band width of the reference amplifier and d.c. amplifier loop, closed, is approximately 0·01 cps. The complete voltage control loop reaches unity gain at 12 cps, and a phase margin of 45° is maintained at all times. The chopper shown in the block diagram is a fairly conventional switching transistor application followed by an emitter follower which drives the filter. The filter produces 0·15 per cent total harmonic distortion at 10 V peak-to-peak square wave input and 2·9 V r.m.s. output. Stability requirements for the chopper, filter and power amplifier are not very severe because of the large gain isolating these circuits from the reference voltage within the reference amplifier. Distortion specifications and voltage gain requirements for the power amplifier have thus been established by the configuration of the voltage control loop.

POWER AMPLIFIERS

While developing the voltage control circuit it became apparent that difficulties would be encountered using silicon controlled rectifiers in the power output stages. Inherently the output distortion of a silicon-controlled rectifier power stage tends to change with load, due to loading of the filter circuits, and due to methods which have to be employed to obtain output voltage regulation. Regulation is obtained by changing the conduction angle of the silicon controlled rectifiers. This changes the distortion components entering the filter. A filter designed to reduce the variable input distortion, while supplying a variable load, to a point where the total change in the output distortion is less than 0·5 per cent, tends to be complex and bulky. Class *B* amplifiers can be designed to meet the distortion requirements, but due to the inherent nonlinearities of the silicon power transistors, power amplifiers with 1 per cent distortion present significant problems. Class *B* amplifiers are also rather inefficient. One of the primary objections to class *B* amplifiers in static inverters arises from inefficiency causing dissipation problems in germanium power transistors. Relatively low maximum heat sink temperatures are therefore dictated if germanium transistor class *B* amplifiers are used. Fortunately silicon power transistors are capable of dissipating considerable power even at 100°C heat sink temperature. In order to decide on a particular power amplifier

to be used, a number of configurations were investigated with the assistance of digital computer facilities. The particular configuration chosen combines reasonable simplicity with low distortion and fair efficiency. The theoretical maximum efficiency of a class *B* amplifier is $\pi/4$ and may be derived as follows:

$$P_0 = E_{\text{r.m.s.}_0} I_{\text{r.m.s.}_0} = \frac{E_0 I_0}{2}$$

$$P_i = E_S \frac{1}{\pi} \int_0^{\pi} I_S \sin\theta \, d\theta = \frac{2}{\pi} E_S I_S$$

$$\text{Efficiency} = \frac{P_0}{P_i} = \frac{\pi}{4} \frac{E_0}{E_S} \simeq 78{\cdot}5\% \frac{E_0}{E_S} \qquad (1)$$

where subscript S refers to d.c. supply quantities, i to input and 0 to output quantities.

The amplifier designed for the static inverter generates a sine wave in three time intervals. The operation will be explained with reference to the block diagram in Fig. 4. A low power output stage conducts from zero degrees of a sine wave cycle until it approaches saturation at an angle θ. For purposes of illustration let θ be 30°. The power output of this low-powered stage is coupled to the load by means of a transformer with a ratio N_1. For this example let N_1 equal one half. At the 30° point the output voltage will therefore approach $E_S/2$. When the lower power amplifier is within a predetermined small voltage from saturation, drive to this stage is switched off, while drive to the main power stage is turned on. The main power stage will approach saturation at 90°. The main output is coupled to the load through a transformer ratio $N_2 = 1$.

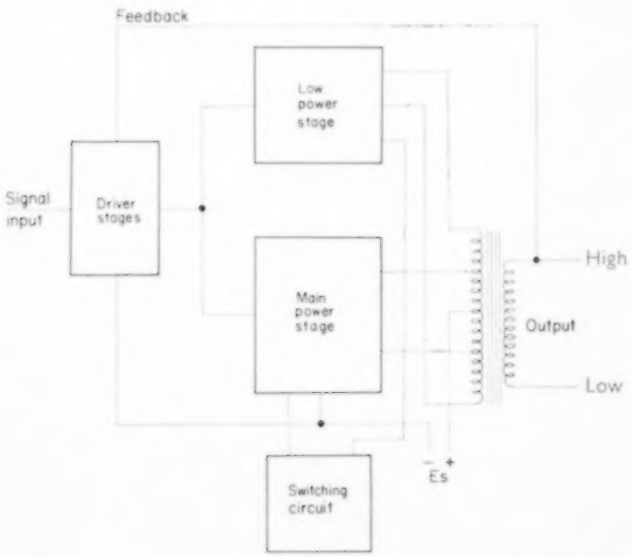


Fig. 4. Block diagram of inverter power amplifier.

In practice saturation resistance, transformer losses and power taken by the driver stages will reduce the theoretical efficiency. Designing for a supply voltage range such as 25–29 V gives a nominal E_S larger than the minimum E_S. This again lowers the efficiency.

At 90° the output voltage is therefore E_S. The process is repeated in reverse from 90°–180°, generating a sine wave at the output terminals. By proper selection of the gains in the two amplifiers, nonlinearities in the output stages can be minimized. An improvement in efficiency

over a class *B* amplifier is achieved and can be computed as follows:

$$P_0 = \frac{E_0 I_0}{2}$$

$$P_i = E_S \frac{2}{\pi} \int_0^{\theta_1} (\sin \theta_1)(I_0 \sin \theta)\, d\theta + \int_{\theta_1}^{\pi/2} I_0 \sin \theta\, d\theta$$

$$\text{Efficiency} = \frac{P_0}{P_i} = \frac{\pi}{4} \frac{E_0}{E_S} \left[\int_0^{\theta_1} (\sin \theta_1)(I_0 \sin \theta)\, d\theta + \int_{\theta_1}^{\pi/2} I_0 \sin \theta\, d\theta \right]^{-1} \quad (2)$$

Fig. 5 presents an evaluation of efficiency made by using equation (2). An additional gain over the class *B* amplifier efficiency occurs at supply voltages other than minimum because the low power amplifier will approach saturation regardless of supply voltage. The maximum swing of this amplifier is not controlled by the output voltage. The efficiency at which the low power amplifier operates is therefore independent of supply voltage and is maintained at maximum.

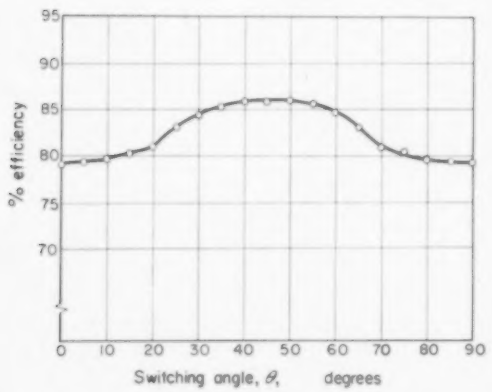


Fig. 5. Theoretical efficiency of switched amplifier at minimum supply voltage.

Short-circuit protection for the main amplifier is obtained automatically because the switching circuit will not supply drive to the main amplifier unless an output voltage of proper magnitude is present. During a short-circuit condition the main amplifiers will, therefore, never be turned on. Additional linearity and phase stability is obtained by the use of negative feedback around the entire amplifier. Approximately 24 dB of negative feedback are used in each of the power amplifiers of the two-phase static inverter. Only sufficient detail to illustrate the concept has been presented here. Much more detail has been evolved in order to properly design the actual unit.

PERFORMANCE CHARACTERISTICS

Performance of the static inverter was evaluated in terms of a single-phase loop and a two-phase complete circuit. The single-phase loop was used primarily to evaluate the action of the control circuit. Regulation and distortion versus load power are plotted in Fig. 6. Voltage regulation within $\pm 0{\cdot}1$ per cent was obtained from zero to full load. Exposure to a temperature range from -20° to $+100^\circ$C resulted in a voltage variation of less than 0·2 per cent. For the temperature range of 0°–85°C at the control unit, with 100°C at the heat sink, results within 0·1 per cent were obtained. The two-phase inverter was built with a single control loop regulating the average of the two-phase voltage as outlined in Fig. 7. This configuration was chosen to eliminate the bulk of two control loops. Due to the load unbalance specification limiting the maximum unbalance to 10 per cent for the particular application, voltage regulation better than 0·3 per cent was obtained for the requirements outlined. Overall efficiency of the two-phase inverter ranges from 67 per cent maximum, at rated conditions and minimum supply voltage, to 60 per cent nominal and 55 per cent minimum at maximum supply voltage. As shown in Fig. 8, output distortion of the smaller amplifier remained relatively constant between 0·4 per cent and 0·6 per cent for the specified load range. The phase *A* amplifier distortion increased as load was decreased.

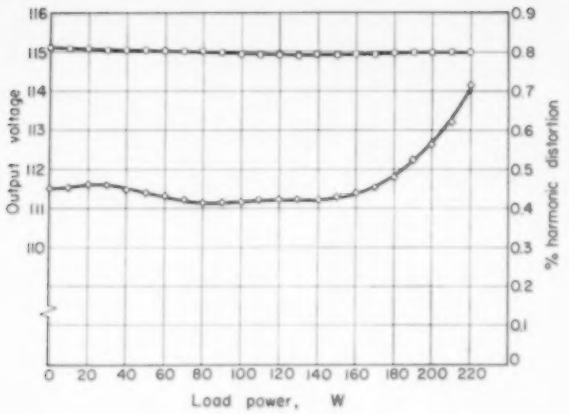


Fig. 6. Single phase loop performance. Per cent total harmonic distortion vs. load power.
○ Output voltage (volts r.m.s.) vs. ◊ load power (watts at 80 per cent p.f.)

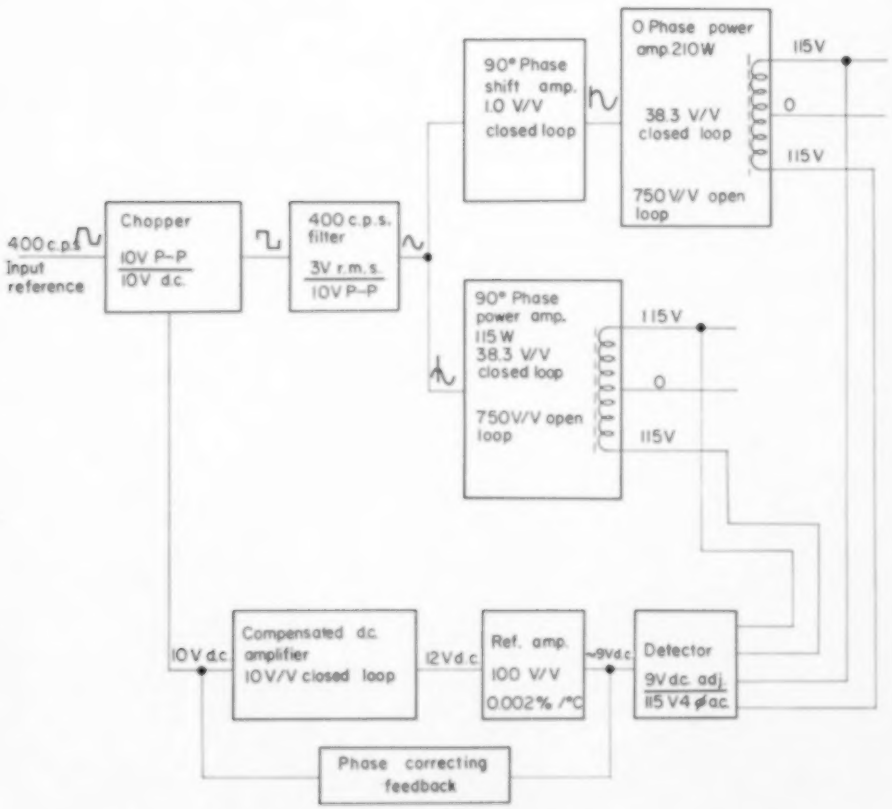


Fig. 7. Two phase 400 c.p.s. static inverter block diagram, single control loop.

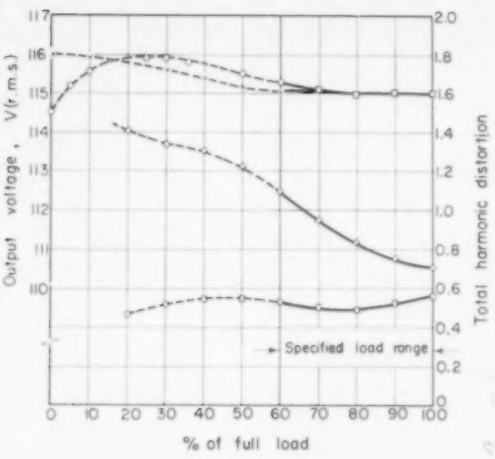


Fig. 8. Performance of two phase static inverter.
○ A phase output voltage ◊ A phase distortion
× B phase output voltage □ B phase distortion

A large power factor correction capacitor which was permanently connected to the amplifier caused overloading of the lower power section of the amplifier as the effective power factor shifted due to load reduction. Individual power factor correction of load sections would eliminate this problem. The effect of supply voltage variations from 25 to 29 V d.c. was negligible. Supply voltage variations also effect the distortion less than 0·1 per cent.

CONCLUSION

The design was undertaken to provide a static inverter capable of supplying precision power while operating at 100°C heat sink temperature, and to test the degree of voltage regulation obtainable. Distortion figures below 1 per cent and regulation to 0·1 per cent can be obtained for specific loads. The voltage control loop outlined in Fig. 1 has been refined and temperature controlled for use in a low power

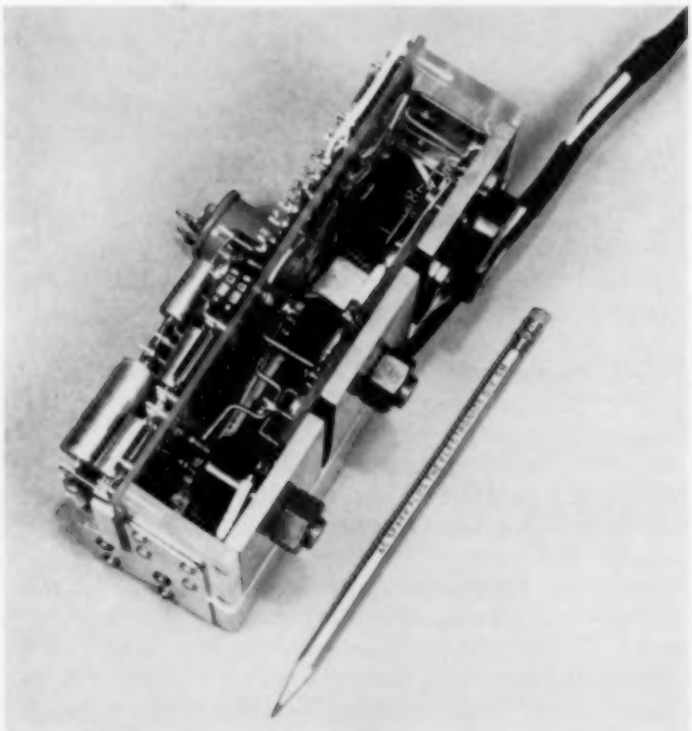


Fig. 9. 220 W single phase static inverter laboratory model, side view.

(10 W) high-precision supply. Output voltage stability better than $\pm 0{\cdot}025$ per cent has been obtained during several hundred-hour test periods. Maximum power per phase for static inverters of the type outlined in this paper is limited by the silicon power transistors available.

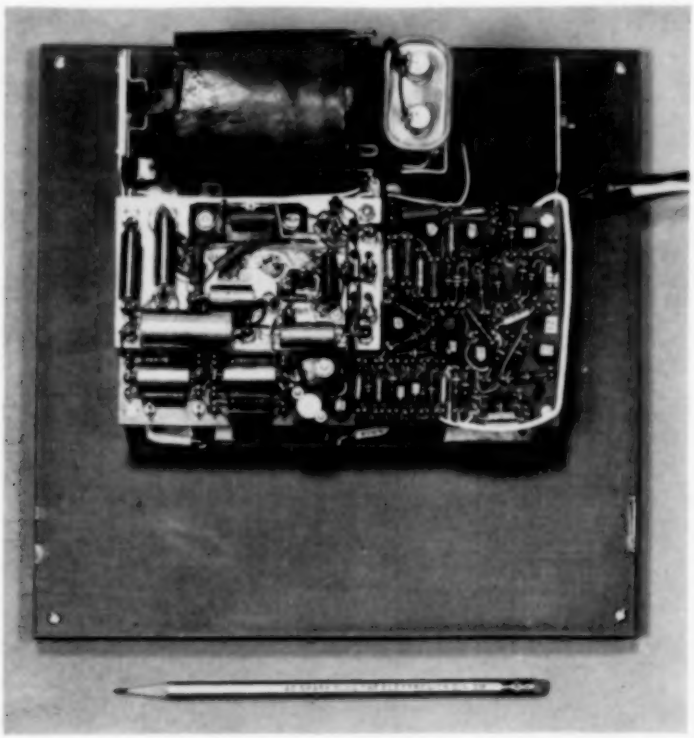


Fig. 10. 200 W single phase static inverter, top view with laboratory heat sink.

50 W per phase can be obtained, without paralleling of transistors and using commercially available units. 20 A units which are scheduled to become available late this year will make phase powers of 200–250 W per phase possible in a very small light weight package. A 225 W amplifier was constructed with prototype 20 A silicon power transistors and was found to be very successful. Very accurate, rugged, light weight, static inverters of up to 600 VA at 70 per cent power factor for two-phase and 1 kVA for three-phase are within the state-of-the-art and will become available as soon as 20 A silicon transistors become available in quantity. The 500 VA two-phase unit constructed with the use of 5 A silicon transistors weighs approximately 17 lb. This weight can be reduced with optimized transformer designs. The unit is somewhat bulky due to the use of parallel output transistors. With the present cost of silicon power transistors, the design is applicable only where precision performance and temperature requirements justify the high cost. As the state-of-the-art advances in silicon power transistors the cost of units will no doubt decrease. The design outlined in this paper promises to be very usable in the smaller guidance systems and similar devices of the future that require power supplies with limited power requirements and greater precision.

A PRACTICAL SOLUTION TO THE ARCING PROBLEM AT HIGH ALTITUDES

BRADLEY R. SPINK
AC Spark Plug Division, General Motors Corp.

Abstract—Malfunctioning of ballistic missile guidance equipment due to arcing at high altitudes is discussed in the light of a reconsideration of original design and design evaluation test techniques. Corrective methods are developed for the existing equipment resulting in sealed components to entrap ground atmosphere during flight. As an additional measure, partial encapsulation processes are used to exclude air from critical areas where voltage, terminal spacing and altitude could combine to cause electrical breakdown.

INTRODUCTION

Early in 1956 AC Spark Plug undertook to design, develop and put into production the inertial guidance system used in the Thor missile. Less than 2 years later during the early stages of flight testing the problems to be expected at this phase of the program developed. A portion of these problems was traced to arcing in the guidance equipment at the higher altitudes. In spite of the considerations given to the altitude problem in the original package design, and in spite of the design verification tests performed in the environmental laboratory, each of the systems manufactured up to that time was a potential failure in flight unless corrective action was taken.

The objective of this paper is to point out the pitfalls which may be encountered when developing equipment for use at altitudes where arcing becomes a problem, and to suggest the remedial measures which might be taken if the problem is already existent.

DESIGN PHILOSOPHY

One of the considerations in the design of the inertial guidance system for the Thor was a review of Paschen's law in order to determine the conditions required for arcing and thus avoid them in the design. Briefly, this law states that the arcing voltage for an electric field is a function of the product of the gas pressure and the distance between bare conductors. Figure 1 is a curve illustrating Paschen's law, showing the breakdown voltage for air in a uniform field.

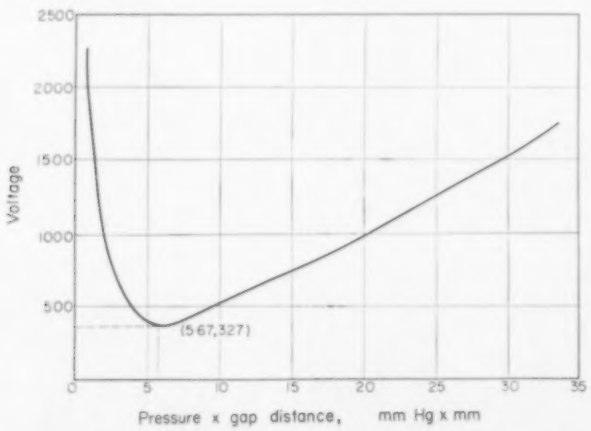


Fig. 1. Breakdown voltage in air-uniform field (approximate).

It can be seen that with a decreasing pressure-gap distance product, occurring as the missile attains altitude, the potential required for arcing also decreases, approaching a minimum at 327 V. With a further increase in altitude, the potential required for arcing rises.

The curve is useful in two ways. (1) It reveals the fact that if the voltages in the equipment are kept low enough, arcing cannot occur at any altitude. (2) If the voltages are above the minimum required for arcing, the critical altitudes can be determined from the known gap distances in the equipment. Assuming that gap distances between 2 mm and 20 mm cover all of the practical situations, the critical region for arcing would then lie roughly between 100,000 and 200,000 ft, a range in which the guidance equipment would be performing its function during the flight of the missile.

Fig. 2. Unpotted power supply showing hard packaging.

No difficulty with arcing was anticipated in the system which was designed. The highest voltage used in the equipment is 270 V d.c. It was also felt that the components or black boxes making up the system would be at least partially sealed simply by the nature of their construction. This would tend to entrap the ground atmosphere and cause a sufficient lag in the lowering of internal pressures to complete the short flight time involved.

The design of the component containers incorporated "hard" packaging in some cases and "channel" packaging in others. Hard packaging is a scheme in which the parts and assemblies are directly mounted on the walls of the box (Fig. 2). In a channel package (Fig. 3) the parts and assemblies are mounted on *H*-shaped aluminum extrusions which in turn are bolted directly to the base casting and cover plates of the container. The connectors and the covers employ flat gaskets for sealing to the container.

Fig. 3. Unpotted stabilization electronics showing channel packaging.

REVEALING THE PROBLEM

Prior to flight testing, altitude tests were performed on assemblies including the maximum equipment possible to fit into the limited space in a bell jar (Fig. 4). A bell jar was used because it was the only facility available at the time which

was capable of simulating altitudes in the critical region mentioned. Although these tests were judged successful because no arcing occurred, the bell jar facility had several disadvantages. Each of the black boxes practically filled the bell jar, creating thermal problems when the equipment was energized for any prolonged period. To solve this problem the components were energized only at discreet intervals of pressure during the test runs. It now seems probable that many critical pressures were passed by where arcing might have taken place had the equipment been energized. Furthermore, since the bell jar had only a small number of external connections, the components could not be completely energized during the tests. Generally, the components were tested with their covers off and with one or more representative circuits energized at normal operating voltages and under static operation. The conclusions drawn from the bell jar tests were that arcing potentials did not exist in the equipment. When it later became apparent that arcing problems at altitude did exist in flight, it was realized that better facilities and test techniques would be needed in order to simulate the flight conditions in the laboratory and to verify any modifications incorporated in the equipment to prevent malfunctions.

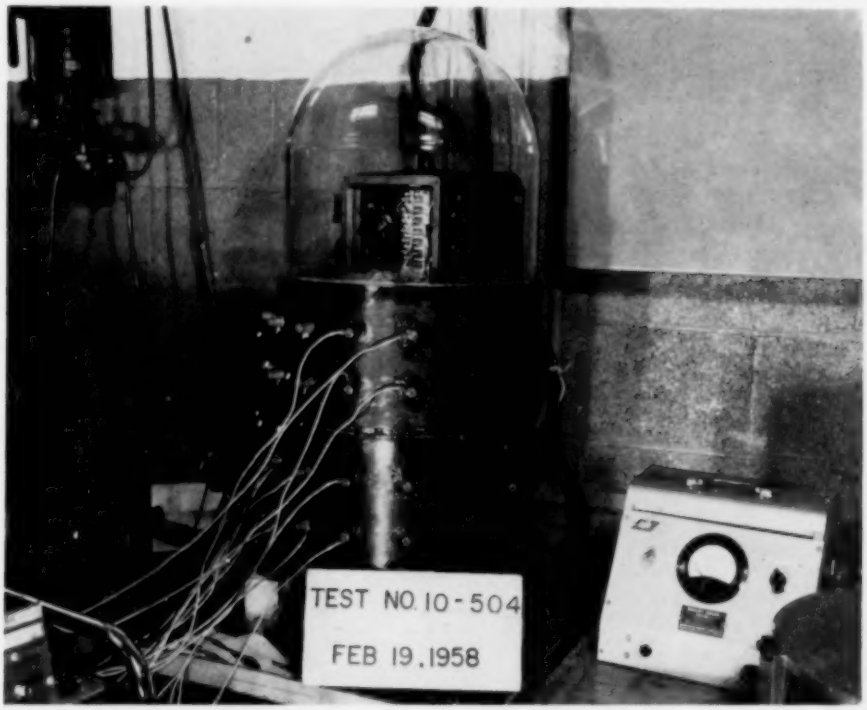


Fig. 4. Bell jar facility.

ATTACKING THE PROBLEM

It was determined that testing of complete systems, fully energized and operating under dynamic conditions, would be the only satisfactory way of proceeding. Consequently a large altitude chamber was prepared for testing equipment up to 200,000 ft (Fig. 5). The first system tests in this chamber indicated that several areas of trouble existed. Some of the places where arcing occurred were between pins on

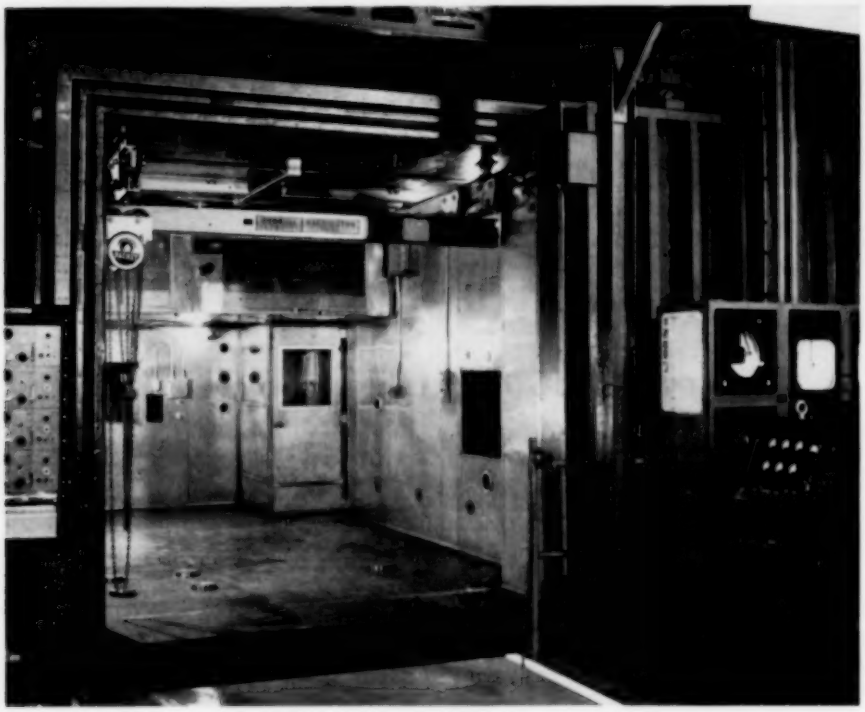


Fig. 5. High altitude chamber.

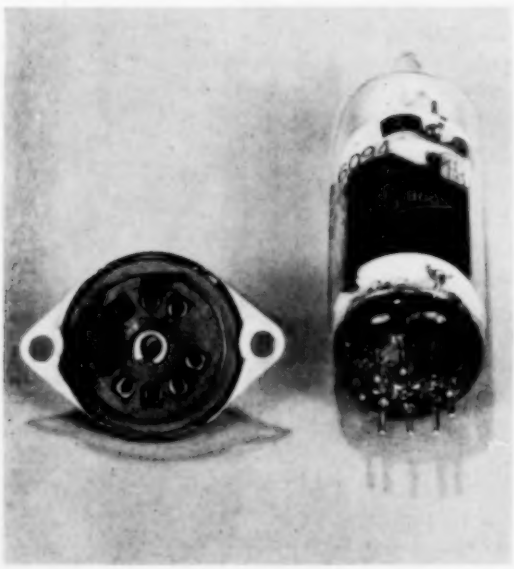


Fig. 6. Arcing between pins on tube.

tubes, at the base of the tube sockets (Fig. 6), between connection points on terminal boards (Fig. 7), and at output transformer and magnetic amplifier terminals.

The next step was to initiate pressure leak tests on all sealed components. These tests were conducted by adding Schroeder valves to the components and then pressurizing the closed components to 2 atm. The return to room pressure was timed. The test criterion was that the leak rate should not exceed 3 p.s.i. in the first 5 min. The initial tests showed that the boxes leaked so fast that the initial 2 atm could never be reached by pumping. These boxes rather than being sealed, reacted much like sieves. If there had been time for a complete package redesign, a simple, positive seal of the containers would have been accomplished by attaching all internal assemblies to the box in a manner to eliminate the need for bolting from the outside, and by avoiding casting shapes where numerous cored

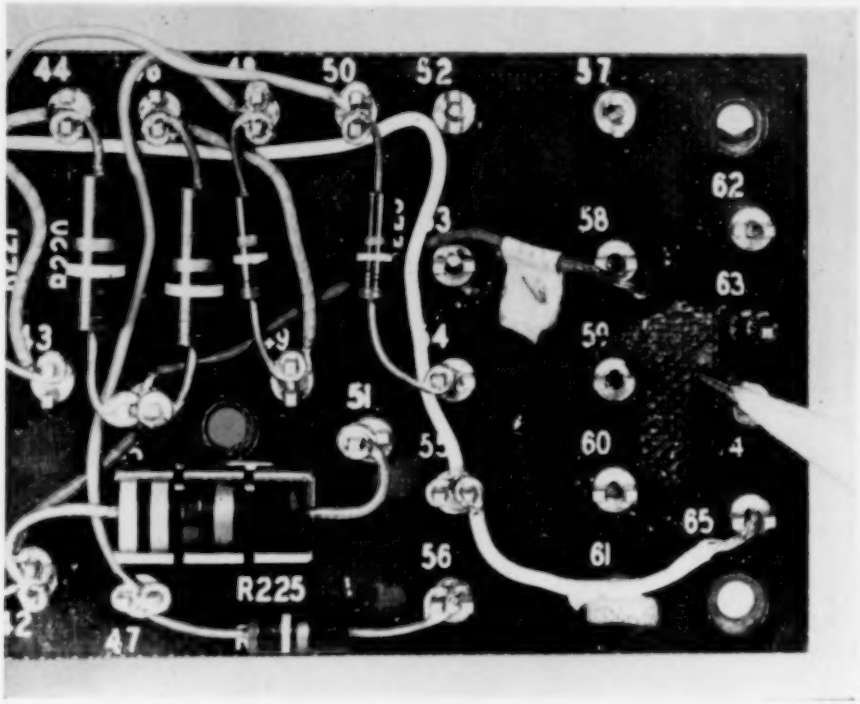


Fig. 7. Arcing on terminal board.

sections in the casting could cause porosity of the metal itself. By these means the sources of possible leaks could be reduced to the covers and connectors. The covers would be sealed with "O" rings embedded in a groove in either of the mating parts. The main advantage of this type of construction over the flat gaskets now employed is a metal-to-metal contact enabling proper torques to be applied to the fastening screws. This would prevent loosening of screws or covers during vibration.

BFR type bulkhead connectors would be used instead of the series *E* connectors presently used. Although both are gas tight, the BFR connector being a thread mounted type would eliminate the need for four attaching screws and thus remove many potential leak spots. These connectors are also equipped with an "O" ring which has proven to be the most effective type of seal.

Since a redesign could not be undertaken and effected in time to meet flight commitments, immediate steps were taken to correct the problems with the existing equipment.

The most critical leaking occurred at several hundred screw holes through the boxes where internal assemblies were mounted, or where connectors were attached. Substituting Tuff-Tite sealing screws, a type of screw utilizing a plastic cap seal, for the original screws provided a suitable fix in this area. It would have been desirable to use "O" rings in place of the flat gaskets on the covers, however, there proved to be an insufficient area on the flange surface to allow a groove to be cut.

Another location at which arcing occurred was at the connector of the torque exciter. This was the one place where a hard connector was in use. This connector was not designed for high altitude use but could not be replaced because the same torque exciter was also used in other projects where this type of connector was required. A manufacturing problem was involved wherein destruction of standardization would

result in an impractical outlay of money. Therefore, it was necessary to develop a method of sealing to provide protection. The final result was to insert a soft rubber gasket over the pins. This gasket was made thick enough to assure making contact against the hard inserts of the connector when mated, thus providing a reasonably air-tight seal.

Subsequent static leak tests indicated that all the modifications made were effective. In most cases no drop in pressure was measurable after several hours had elapsed. Even vibration tests had little effect on increasing the leak rates of the pressurized boxes.

Simulated altitude tests with the boxes sealed and the equipment operating as in flight, proved that protection against electrical breakdown had been accomplished.

BACKUP FOR RELIABILITY

Because of the high reliability demands of the missile program it was decided not to depend entirely upon the seal for protection. Considering the hundreds of places where corrections had been made, it was reasonable to take the attitude that some modified area might suddenly fail at an inopportune time. Transportation and storage conditions such as temperature and vibration in addition to flight environments, could conceivably destroy the effectiveness of a seal at high altitudes even though it was checked just prior to a flight.

Since the sealing methods incorporated on the equipment boxes could not be depended upon absolutely to maintain internal pressures at a level outside the critical range for arcing, a backup program of covering was introduced. This device is based on the fact that if air is completely excluded or kept from impinging directly on bare conducting surfaces, arcing cannot occur. That the two preventative methods—sealing to entrap air and covering to exclude air—could be employed simultaneously, served greatly to enhance reliability.

The obvious way of excluding air would be simply to encapsulate the entire equipment. However, such encapsulation was impractical because the added weight of complete encapsulation could not be tolerated, and because repair and maintenance of the equipment would be seriously impaired

In view of these considerations the "thin coat cover" technique was adopted and applied only to those areas which were considered critical to the arcing problem. A complete wiring analysis was made to determine where potential differences of 275 V or more could occur at gap distances of 1 in. or less. That 275 V was taken as a standard at this point instead of the 327 V indicated by Paschen's law, provided a reasonable margin of safety, since it is known that the ideal conditions assumed for Paschen's law to hold do not exist in actual practice. By knowing these critical areas and covering them, the necessary insulation incident to the wiring throughout all the connection points in the equipment was accomplished and gave the effect of total encapsulation without the accompanying disadvantages of that process.

Many critical locations were revealed. In practically all cases the high potential differences discovered were created by the 270 V d.c. at one terminal and an a.c. voltage at the other. As an example, the 270 V d.c. adjacent to the system 115 V. a.c. on a terminal board would produce a potential difference of over 430 V on the negative swing of the a.c. voltage. In other cases, signals, while not presenting a problem during static operation, were known to reach high enough levels during dynamic operation to constitute a definite altitude hazard when routed adjacent to the 270 V d.c.

The choice of covering material to be used on these areas was dictated by several factors. Where several terminals were involved in a concentrated area, limited encapsulation was made over that area even though it meant covering some terminals that were not a part of the problem (Fig. 8). These areas were potted with Silastic RTV-501, a rubberlike material that could be cast into the desired shape. This was generally accomplished by removing the sleeving from the terminals, damming the local area with a plastic material, sealing with tape, and filling the dam until the terminals and the exposed leads were covered by at least $\frac{1}{8}$ in. of

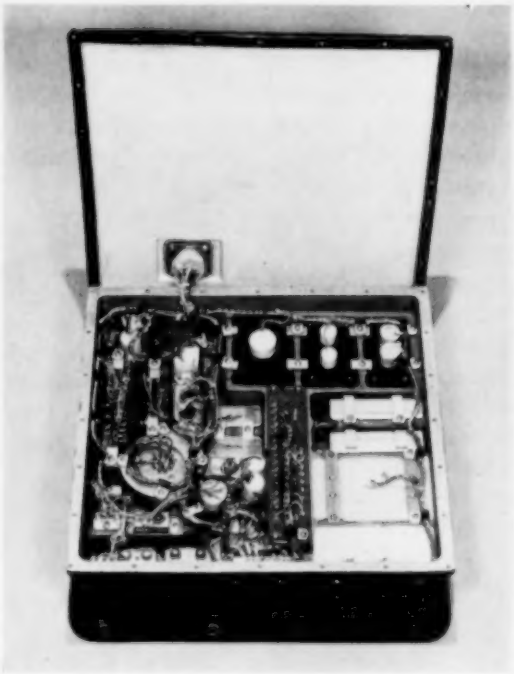


Fig. 8. Partial encapsulation of power supply.

the potting material. Approximately 8 hr at room temperature is required for this material to set before the dams can be removed. In cases where only one terminal was to be covered, the most effective material for this purpose was found to be a 1:1 mixture of Silastic RTV-5302 and -5303 applied so as to cover the terminal and exposed lead with a $\frac{1}{8}$ in. coating (Fig. 9). In handling the problem with tubes where the capability of removal had to be retained, a silicon grease was applied $\frac{1}{8}$ in. thick over the entire tube base so that when the tubes were reinserted there would be no chance for air to reach the pins (Fig. 10).

That these covering techniques alone would have proven effective has been demonstrated repeatedly in the altitude facility where tests were run successfully with the boxes open. While other cover material might have been found to be equally effective, these had characteristics which were particularly suited to the problem. Primarily they covered well, in contrast to other materials such as varnishes and shellacs which had a tendency to flow away from sharp edges and points no matter how many coats were applied. The fact that the material finally used would set at room temperature was a decided advantage over some that required oven curing at high temperatures, and in the process caused damage to insulation and assemblies. Finally, although adhering well, these materials are easily removed should repair of the equipment become necessary. Approximately 165 separate areas required covering. It is needless to say that good workmanship practices must be maintained if these procedures are to be depended upon as a reliable backup for sealing. If any of these areas were to be inadvertently left uncovered, either completely or partially, a catastrophic failure would almost certainly result in flight should the component seal become ineffective.

CONCLUSION

To sum up, it should be remembered that if

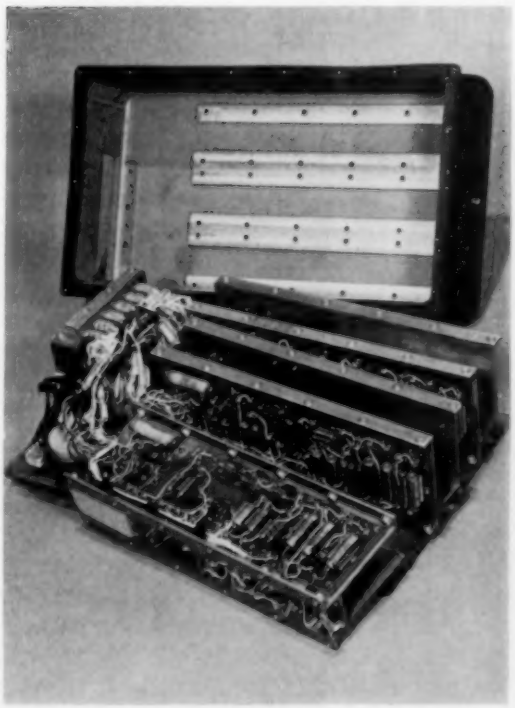


Fig. 9. Partial encapsulation of stabilization electronics.

steps are taken in the beginning to assure positive and reliable sealing of black boxes there is no further need to be concerned with encapsulation or partial potting techniques. However, in a situation where protective measures are taken at a later stage, when there is no possibility of redesign and where there is no margin for error, it is well to adopt the double dependency measures of potting and sealing, to insure that the equipment will operate without arcing in the high-altitude regions.

Since taking these protective measures on the Thor guidance equipment, subsequent flights and altitude tests have verified that the corrective methods were positively effective. The basic principle of sealing boxes and covering marginal areas where voltages, terminal spacing and altitude can combine to cause electrical breakdowns has been followed through to a successful conclusion.

Fig. 10. Protection for tubes with silicon grease.

INERTIAL COMPONENTS TO SUPPLEMENT ASTRO-FIXES IN SPACE GUIDANCE

CHARLES J. MUNDO*
Airborne Equipment Division, Raytheon Company

Abstract—In this paper the requirements for inertial sensors in space navigation systems are discussed. Classically, terrestrial navigation has been based upon direct sensing of position and coordinate orientation. However, in space navigation another set of variables—thrust acceleration components—must be added to the conventional terrestrial set of measurements. These new variables are needed because, (i) different from terrestrial flight, an increment of applied thrust does not merely affect the course of the vehicle at the time it is applied, but it alters the whole path to the target and thus, if erroneously applied, must be totally cancelled out, (ii) the measurements of position are so coarse that a very large error in position must be accumulated before a correction can be instituted.

Inertial elements serve to ascertain instantaneously that trajectory-correcting thrust is being applied in the correct direction and that the total impulse is correct. However, the inertial components have no capacity for sensing the errors which arise during ballistic flight due to (1) uncertainties in astronomical units, (2) uncertainties in gravity, (3) integration errors internal to the system.

The position elements may be viewed as the low-frequency sensing elements which measure the accumulated errors, while the inertial elements are viewed as the high-frequency sensor which measures instantaneous errors in the system.

In synthesizing a guidance system, a cross-over filter must be designed to accept information from each system. The nature of the cross-over filter depends upon the mission, the optimization of fuel consumption during the flight and the thrust. An example is studied involving a flight to a near planet with a 1000-mile unit terminal accuracy, including the effects of various thrust profiles upon component accuracy. Effects such as accelerometer bias errors, scale errors, nonlinearities and gyro errors are presented in terms of error at the target as a function of thrust profile. These curves are used to derive a cross-over filter accepting both inertial and position-sensed data.

This paper is concerned with the accuracy problems that will arise in the design of guidance equipment for initiating a ballistic flight from the Earth to Mars, and with low dynamic smoothing can be utilized to cope with these problems. More specifically, this paper is concerned with the nature of the errors in the guidance equipment itself and not with the optimal choice of a trajectory for such a flight. This problem offers unusual difficulties in two respects. (1) Astronomical units which are used in describing the motion of the planets are uncertain in their relationship to Earth units to 1 part in 2000. This alone would result in a miss distance of 40 Mars diameters if the launch velocity were measured in Earth units.[1] (2) Gravitational uncertainties of the Earth–Moon system, if terminal conditions were established in this region, would result in a miss distance of 12 Mars diameters.

To overcome these two problems, the guidance systems for interplanetary flight must measure the dynamic variables of flight for cutoff in astronomical units and at a distance sufficiently removed from the Earth–Moon system so that the gravitational effects of the Earth are of little consequence.

The only system which measures in astronomical units is a celestial system. In addition to its capacity to measure in astronomical units, the astro system is self-contained, does not depend upon its adjacency to the Earth for measurement, and does not involve the complexity of a rapidly rotating Earth-fixed base line.

A position measuring system alone is not sufficient to attain the required precision of guidance. Therefore, it is assumed that an inertial system[2] which measures the guidance correction acceleration is used to supplement the

* Formerly with Arma Division, American Bosch Arma Corporation.

celestial equipment. The inertial system measures small changes or is calibrated in trajectory by the celestial components and, therefore, its calibration tolerances are not of major concern.

It can be assumed that the guidance regimen is divided into two phases. The first is concerned with the escape of the vehicle from the gravity field of the Earth and, for reasons already pointed out, full guidance cannot be accomplished accurately in this phase. In the second phase, after escape from the Earth–Moon field, a second impulse is used to add the final trim in velocity accuracy that eliminates the probable miss of 40 Mars diameters inherent in the system at the termination of the first guidance phase. This second guidance phase is required to measure the velocity to 1 part in 10^5.

The celestial navigator determines the navigational parameters of velocity and position by establishing the time sequence of positions through which the vehicle has passed. The relative angular position of three planets is measured and from known planetary constants the position of the vehicle is computed. With current techniques of tracking, stabilization and angular pickoffs, results accurate to 10 sec of arc can be attained. This can probably, in the near future, be improved by a factor of 100 if star background comparison[3] is used to locate the planets angularly. Thus, individual measurements of vehicle position accurate certainly to 4000 miles and probably to 40 miles can be established based upon the near planets. A substantial sequence of such readings, along with the time when they are taken, are obviously necessary to establish the trajectory with sufficient precision to meet the guidance requirement for a collision course to Mars.

To state this in another way, a lag exists between the time when the data are first sensed and the time when the position and velocity of the vehicle are known with sufficient accuracy to institute a correction in the velocity of the vehicle. The increase in the quality of the guidance data as a function of smoothing time is shown in Fig. 1. These curves can readily be derived by assuming a Gaussian distribution in the errors of measurement and then fitting the best parabola through the data by the method of moments. Thus, the r.m.s. error in the guidance measurement will be

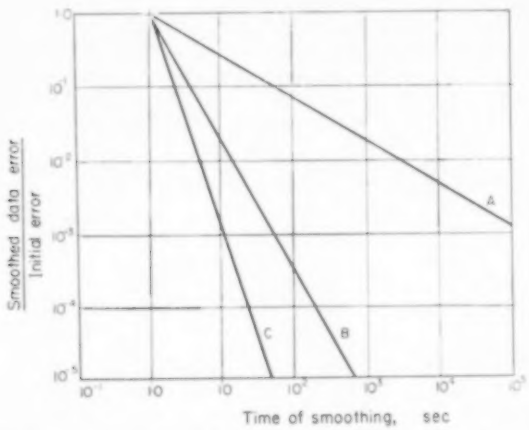


Fig. 1. Curve A, position uncertainty; curve B, velocity uncertainty; curve C, acceleration uncertainty.

$$\sigma_x = 0{\cdot}75\delta\sqrt{(a/\tau)}$$

$$\sigma_{\dot{x}} = 0{\cdot}86\delta\sqrt{(a/\tau^3)}$$

$$\sigma_{\ddot{x}} = 1{\cdot}38\delta\sqrt{(a/\tau^5)}$$

where δ = r.m.s. error in position measurements, a = interval between readings, τ = smoothing line.

This situation is only true when the vehicle is

dynamically under steady-state conditions, that is, when no thrust is being supplied to change its energy state. Then such smooth-data celestial techniques as have been described will serve amply to affirm the trajectory through which the vehicle is traveling. However, when the trajectory is in the process of being corrected, the smoothing lag required to ascertain when a sufficient correction impulse is obtained will give rise to one of two problems. The thrust level required to obtain the necessary impulse will be limited to such a low value that the fuel

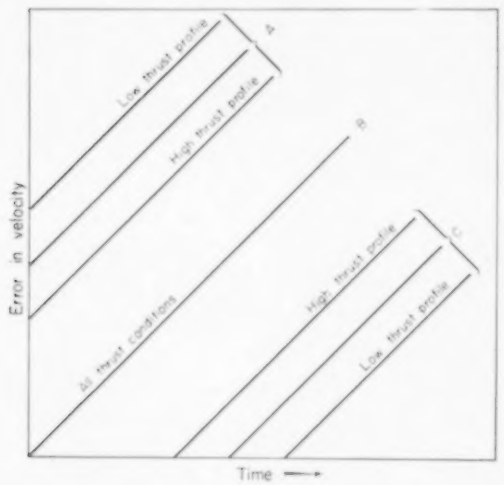


Fig. 2. *A*, bias letters; *B*, scale errors; *C*, linearity errors.

expenditure required to make the correction will be materially increased or the error uncertainty will not fall within the tolerances of accuracy required for guidance. This can be seen from Fig. 2. The error in celestial measurement of velocity (curve *A*)—the critical parameter in guidance—decreases with smoothing time. However, during the smoothing interval, the thrust alters the velocity of the vehicle at a rate proportional to the thrust magnitude. Thus, obviously, for low thrusts, the rate of increase of position uncertainty due to the low thrust (curve *B*) is lower than that for higher thrust (curve *C*), so that a longer smoothing time can be utilized in the low thrust case because of the longer time it takes for dynamic errors to exceed the threshold of the smoothing errors.

The primary attribute of inertial sensors in space guidance is the capacity to instantaneously sense changes in the dynamic state of a vehicle's motion. Conversely, position sensors must accumulate data for a long smoothing interval before an accurate measurement can be made of the change in dynamic state. Unfortunately, inertial sensors cannot directly measure position and velocity in space navigation but must extrapolate these measurements from an accumulation of acceleration history of the vehicle. As a result of extrapolation errors, plus the change in scale factors between the Earth and interplanetary space, the accumulated errors eliminate the possibility of using pure inertial systems for precision guidance. Thus steady-state measurements of velocity and position must be established, by direct measurement such as with the star tracker, while the high-frequency information can be derived from an inertial extrapolator. The question then arises of how to design a filter which will optimally combine the information from the two sources.

The nature of the filter, on the one hand, is dependent upon the errors of the inertial navigator; and these errors, in turn, depend upon the type of degradation in performance caused by the thrust profile on the various components.

The accelerometers of an inertial navigator have three types of errors—bias, scale and linearity. All of these factors combine to cause errors which increase with the time duration of applied thrust (see Fig. 3). However, the thrust profile has a materially different effect on the rate at which these errors are generated. Curves *A* of Fig. 3 show that, as the thrust increases, the bias error expected from the accelerometer diminishes. On the other hand, the scale errors of the accelerometer are independent of the applied thrust as shown in curve *B*, while the linearity errors increase with applied thrust as shown in curve *C*. The optimum thrust profile for the accelerometer then is obtained when the thrust coefficient for the bias error is equal to the thrust coefficient for the linearity factors.

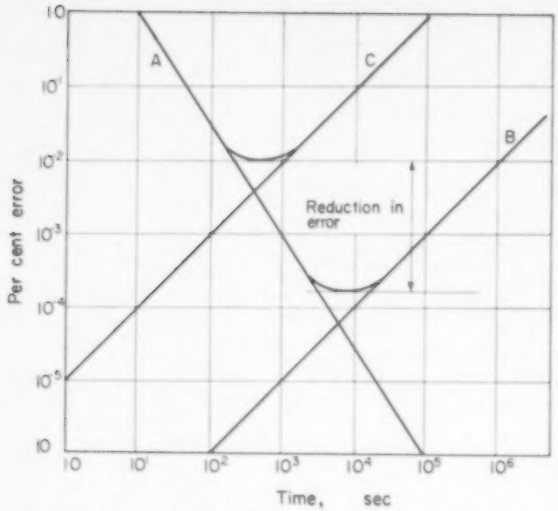


Fig. 3. A, smoothed velocity error; B, low thrust uncertainty; C, high thrust uncertainty.

In this type of system, because of the long effect as an accelerometer scaling error of $2{\cdot}5 \times 10^{-3}$.

The dynamic smoothing of celestial data involves mixing the celestial data with the inertial data in order to formulate the best current estimate of the dynamic variable of navigation for the autopilot to utilize in generating steering and cutoff orders. Fig. 4 shows a filter which is used to smooth astrotracker-derived position data. The transfer of this data to position information is shown in Fig. 4(b). Note how this filter accepts only the low-frequency position data and rejects the high-frequency. The dynamic filter operates on the inertial data in quite a different manner. In this case, the filter serves not only as a filter which passes the high frequencies and rejects the low

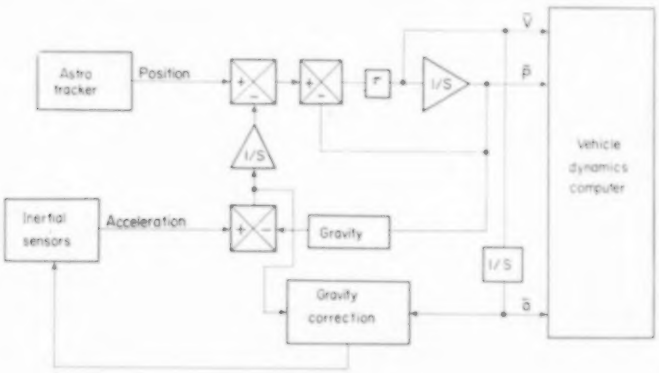


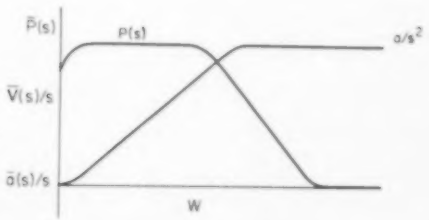
Fig. 4(a).

duration of smoothing, the gyros present a problem which can be overcome by slaving the inertial space reference platform to star coordinates. This problem centers around the free drift of the gyros and is equal to

$$V_{GD} = \tfrac{1}{2}\, a a_0 t^2$$

where $a_0 t$ = drift of gyro in mrad/sec, a = applied acceleration in ft/sec^2, t = time in sec.

For drifts of 2 mrad/hr and smoothing times of 1 day, this error will be of the order of 10 Mars diameters. If the platform is star-slaved to an accuracy of 1 min of arc, this is no longer a serious problem. The error in the slaving alignment of the table will have the same system



$$\bar{P}(s) = \frac{\tau s P(s) + a(s)}{s^2 + \tau s + K}$$

$$\bar{V}(s) = \frac{\tau s^2 P(s) + s a(s)}{s^2 + \tau s + K}$$

$$\bar{a}(s) = \frac{\tau s^3 P(s) + s^2 a(s)}{s^2 + \tau s + K}$$

Fig. 4(b).

frequencies, but also as a short-time extrapolator which derives the gravitational accelerations to establish its effect in the system in the short-term position computation. Further, the space navigator must recalibrate the scale of the accelerometer, which was initially calibrated in the ft/sec² of Earth rather than in the ft/sec² of astronomical space. This is accomplished by the gravity correction computer which slowly modifies accelerometer scale based upon the differences between the smoothed value of acceleration and the sum of sensed and computed accelerations.

The operation of the filter can perhaps be seen a little more clearly if accumulated errors are regarded as a function of smoothing time in the time plane rather than in the frequency plane. Fig. 5 shows a plot of terminal errors vs. smoothing time. Since the terminal error is much more sensitive to velocity than position, the positional errors can be neglected completely. Curves *A* and *B* represent, respectively, a 4000-mile and a 40-mile error in the position fixes, where it is assumed that 100 sec are required to make a complete position fix.

As the smoothing lag increases, the noise constantly decreases, provided that the vehicle is dynamically in a steady-state condition during the smoothing interval. This limitation can be lifted if inertial elements are included in the system to sense and compensate for the dynamic changes. On a constant thrust profile, inertial error increases linearly with time, so that for any configuration of celestial–inertial equipment there is an optimum smoothing time determined by the cross-over from domination by celestial error to domination by inertial error. If the filter cross-over is placed at this point, this particular system is optimized.

In considering the general question of optimization of the filter design, in addition to the inertial system itself, the thrust profile plays an important part because the accelerometer errors are closely related to the magnitude of thrust which they must measure. This places two bounds

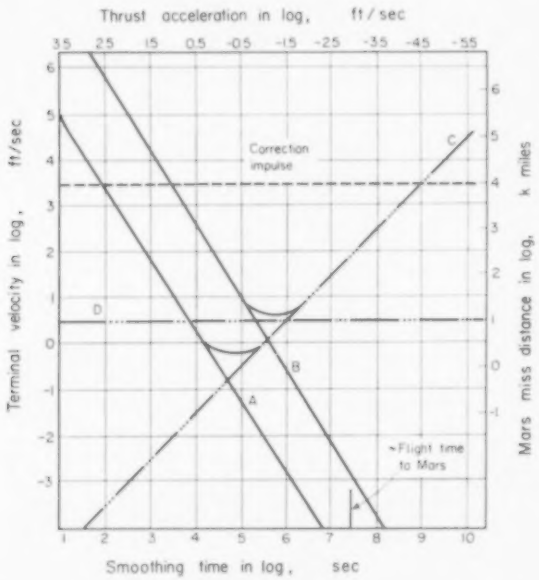

Thrust acceleration in log, ft/sec
3.5 2.5 1.5 0.5 -0.5 -1.5 -2.5 -3.5 -4.5 -5.5
Terminal velocity in log, ft/sec
Mars miss distance in log, k miles
Correction impulse
C
D
B
A
Flight time to Mars
1 2 3 4 5 6 7 8 9 10
Smoothing time in log, sec

Fig. 5. Curve *A*, velocity uncertainty for $S = 40$ miles; Curve *B*, velocity uncertainty for $S = 4000$ miles; Curve *C*, bias error $= 10^{-6}$ ft/sec; *D*, scale error $= 10^{-3}$.

upon the thrust profile—a lower and an upper. In the lower bound, the applied thrust must be sufficiently high so that the correction is accomplished in at least the first half of the flight so that the fuel requirements are not excessive. In the upper bound, the applied thrust must be sufficiently low so that the correction is not accomplished before the celestial data has had time to pass through the filter. Over and above these bounds, there are others which will materially affect the system performance for different types of accelerometer errors. A bias or threshold type of error will result in a fixed level of performance independent of thrust profile for the system, determined by the cross-over point of the two noise curves. With a linearity error, however, the lower the applied thrust the higher the performance level.

The impact of thrust on the pure inertial velocity error can best be demonstrated from the error function of the accelerometer, as follows:

$$\epsilon_V = (K_0 + K_1 a + K_2 a^2)t$$

If a certain specified velocity correction is to be made, the time required to make it will be proportional to the reciprocal of thrust or acceleration, so that

$$\epsilon_V = K_0/a + K_1 + K_2 a$$

It is apparent, therefore, as shown in Fig. 5, that the low thrust profile will result in a larger bias error, that the scale errors will be independent of the thrust profile, and that the linearity errors will be favored by the low thrust profile.

In Fig. 5, lag smoothing time and lag reciprocal acceleration for the set of thrust profiles are both plotted as abcissas; the scales are such that the production of acceleration and smoothing time are equal to the anticipated value of velocity correction for the flight. The improvement of the velocity data derived from position measurements is shown in curves *A* and *B* representing the position fixes of 40 and 4000 miles, respectively. The different types of accelerometer errors are also plotted against the applied acceleration used to reduce these errors. Curve *C* represents a 0·1% scale error and curve *D* represents a 10^{-6} ft/sec bias error. There are two parameters which can be varied to improve the performance of the dynamic filter system—thrust and smoothing time. In the case of the scale error, if the thrust duration T is extended, the error at termination remains constant, 0·001 V_c (where V_c equals the velocity correction). At any time that t is less than T, the accumulated error is

$$\epsilon = \frac{t}{T} 0{\cdot}001\ V_c$$

If t is the chosen smoothing time, the gross error is $\epsilon_{(\text{scale})}$:

$$\epsilon^2_{(\text{scale})} = \delta^2 \frac{a}{t^3} + 0{\cdot}001^2 \left(\frac{t}{T}\right)^2 V^2$$

This $\epsilon_{(\text{scale})}$ can be minimized for a given thrust duration T:

$$\epsilon_{(\text{scale})} = K/T^{3.5}$$

Consequently, if the thrust duration is increased, the magnitude of the system error can be increased indefinitely for a scale error. The same is true for a nonlinear error.

However, this is not so for an accelerometer bias error. The accelerometer bias error at T will be $10^{-6}T$ ft/sec and $10^{-6}T\,t/T = 10^{-6}t$ for any time t. The gross error for a smoothing time t will be

$$\epsilon^2_{(\text{bias})} = \frac{\delta^2 a}{t^3} + 10^{-12}t^2$$

Minimizing t, it is found that t is minimum for

$$t = \left(\frac{\delta^2 a}{2 \times 10^{12}}\right)^{5/2}$$

which is independent of the applied thrust duration, so that no optimization can be made for a bias error by extending the duration of thrust.

The improvement of scale error has particular significance in the present case because one of the difficulties in launching a Mars flight is the inability to calibrate an accelerometer in astronomical units, which now can be accomplished by merely extending the duration of powered flight.

If the case where the astrotracker can measure to an accuracy of 40 miles is taken and the scale

accuracy is assumed initially equal to 0·1 per cent, using a thrust duration of 4 days, an initial error of 3000 ft/sec and a filtering time of 3×10^4 sec, the system error can be reduced to one part in 10^6 of total velocity. This is more than sufficient for a Mars flight provided no outside perturbations are present.

If the case where the astrotracker can measure position to an accuracy of only 4000 miles is taken, the thrust duration would have to approach the duration of the flight before a correction could be accomplished. While some improvement can be achieved in the velocity measurements in this case, it is doubtful that a sufficient amount can be realized to make an adequate correction at an early stage of a ballistic flight.

REFERENCES

1. KRAFFT A. EHRICKE, *Instrumented Comets; Astronautics of Solar and Planetary Probes*, 8th International Astronautical Congress (1947).
2. C. J. MUNDO, *Aero–Space Engineering* **18**, 31 (June 1959).
3. C. D. BOCK. This volume, p. 57.

A METHOD OF DAMPING MISSILE STRUCTURAL BENDING MODES FOR AUTOPILOT DESIGN

LAWRENCE E. FREED and KENNETH L. MILLER
Space Technology Laboratories, Inc.

Abstract—Experience in missile autopilot design has shown that very lightly damped structural bending resonances can be a limiting factor in obtaining satisfactory stability margins. Conventional linear filter techniques are inadequate when the frequency spread between bending resonances and the required pass band frequency is small. An analysis is presented of a proposed method of alleviating this problem by injecting damping into the bending modes through an auxiliary feedback loop. This is achieved by the addition of a small thrust device and rate gyro to the missile. General conditions required for the successful operation of the method are given.

I. INTRODUCTION

Recent experience has shown that missile structural elasticity effects can be a limiting factor in obtaining satisfactory stability margins in an autopilot design.[1] It is not always practical to attenuate the effects of lightly damped (1 per cent or less typical) elastic dynamics on the autopilot by means of linear filters,[2] the determining factor being the frequency spread between the bending resonances and the autopilot band pass frequency. When the spread is small, adequate attenuation at the bending frequencies produces excessive phase lags at lower frequencies which tend to compromise low frequency stability margins and response. Non-linear filters have been attempted without success at this time.[3] In general, non-linear filters imply a system limit cycle; hence, this technique is less desirable than stabilization with linear devices in this respect since a stable linear system is at rest.

The method presented in this paper introduces a reasonable amount of linear damping into the structural modes of oscillation by means of an auxiliary feedback loop. This is accomplished by the addition of a small thrust device and a rate gyro located at the same position on the missile. A swivelling thrust nozzle (or alternatively a thrust device acting normal to the missile axis) is controlled by the gyro sensing the missile rate of rotation. It will be shown that this auxiliary loop can generally be made stable and that the damping in the structural modes can generally be increased. The scheme presented does not depend critically on the bending resonance frequencies nor on the mode shape amplitude, slope or node location.

II. DISCUSSION

1. General System Properties

The basic functional block diagram of the missile autopilot system considered is shown in Fig. 1. The actual design of an autopilot is given in Ref. (1). The same data was used in this study and the results given here are directly comparable to Ref. (1). The conventional autopilot such as that described in Ref. (1) is obtained by considering the auxiliary loop unconnected, $\delta_v = 0$ in Fig. 1 (actually the auxiliary thrust T_v must be made zero also to prevent inertial coupling effects with the main engine).

System stability is determined by considering the Bode diagram formed from the transfer function δ_i/δ_c, the "autopilot open loop transfer function". The effect of the auxiliary damping loop evidences itself mainly by the location in the complex S plane of the poles of δ_i/δ_c. The function δ_i/δ_c is obtained by the solution of a system of $N-1$ linear differential equations (transformed into the S domain by means of the Laplace transformation) in N unknowns. The equations describe the dynamical effects that determine missile motion in a plane. Included are the effects of structural elasticity,

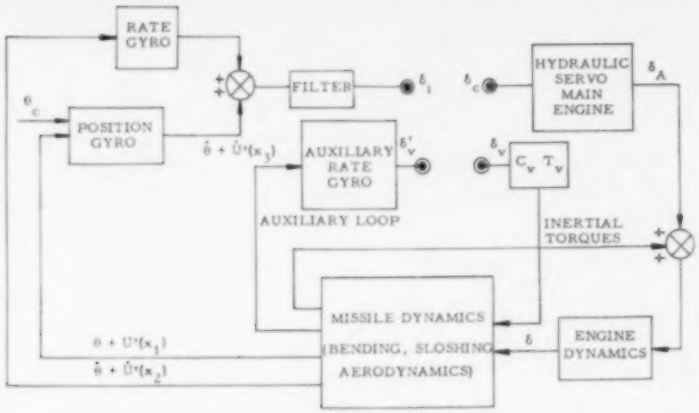


Fig. 1. Basic block diagram.

aerodynamics, fluid sloshing and the dynamics associated with the hydraulic servo-engine systems. A detailed description of the system equations is given in Ref. (4), (5). Modifications of these equations and methods of solution appropriate to this problem are discussed in the appendix.

The nature of the stability problem is seen from the Bode diagrams for δ_i/δ_c of the original system. Plots for two times of flight, in the region of maximum aerodynamic pressure and in the region of burnout, are reproduced from Refs. (1) on Figs. 2 and 3. The sharp resonant peaks are due to the small damping, 0·5 per cent, assumed in the structural modes. Subsequent tests have shown this value to be reasonable. Table 1 lists the damping associated with the resonances of δ_i/δ_c in the original system.

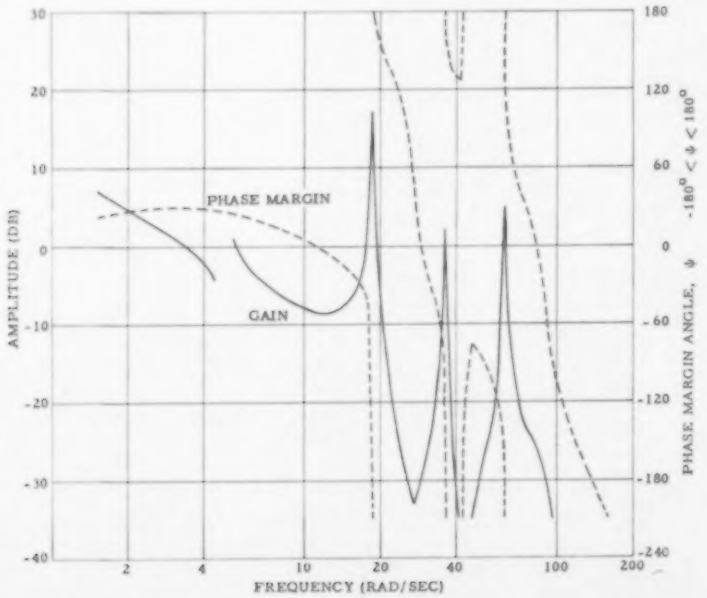


Fig. 2. Bode plot δ_i/δ_c, no auxiliary loop, maximum aerodynamic pressure condition.

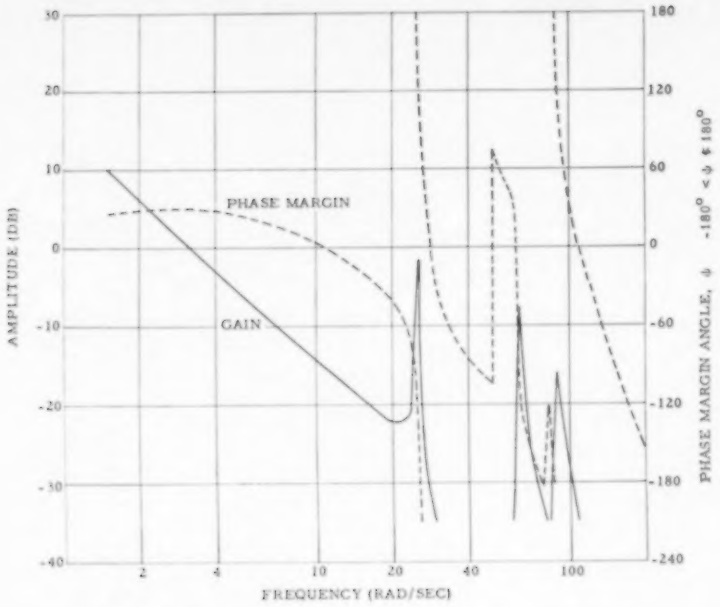


Fig. 3. Bode plot δ_i/δ_c, no auxiliary loop, burnout condition.

Table 1. Bending Pole Location in S Plane, from Ref. (1)

Predominant physical effect	Maximum aerodynamic pressure condition	Burnout condition
1st bending mode	$-0{\cdot}123 \pm j18{\cdot}54$, $\zeta = 0{\cdot}66\%$	$-0{\cdot}098 \pm j25{\cdot}23$, $\zeta = 0{\cdot}39\%$
2nd bending mode	$-0{\cdot}177 \pm j36{\cdot}0$, $\zeta = 0{\cdot}49\%$	$-0{\cdot}594 \pm j64{\cdot}22$, $\zeta = 0{\cdot}91\%$
3rd bending mode	$-0{\cdot}374 \pm j62{\cdot}62$, $\zeta = 0{\cdot}60\%$	$-1{\cdot}84 \pm j88{\cdot}28$, $\zeta = 2{\cdot}1\%$

These values differ from the 0·5 per cent structural damping because of coupling between bending and other dynamical effects. The resonant peaks in the Bode plot will result in system instability if they are greater than unity magnitude (zero dB) and have 180° phase shift. It is not desirable to depend on controlling the phase shift characteristic at the higher bending frequencies since it is difficult to predict all the higher order dynamical effects that may influence the system phase shift significantly in the frequency range of interest.

The proposal advanced here is designed to obviate or at least alleviate the problem of resonant peaking by inserting damping into the bending modes. This is done by sensing missile angular rate and controlling a swivelled thrust nozzle by the angular rate. The rate gyro and thrust controller should be located at the same station on the missile. Usually the missile nose or tail will be a suitable location from a theoretical point of view. An analysis of the efficacy of this arrangement follows.

2. Stability of Auxiliary Loop

Consider the auxiliary loop only and take $\delta_c = 0$. Then the open auxiliary loop transfer function, δ_v'/δ_v, is formed as follows. The rate gyro output is written as (see Fig. 4)

$$\delta_v' = s\theta + sU_v'. \tag{2.1}$$

Adopting the notation of Ref. (4), θ is the rigid

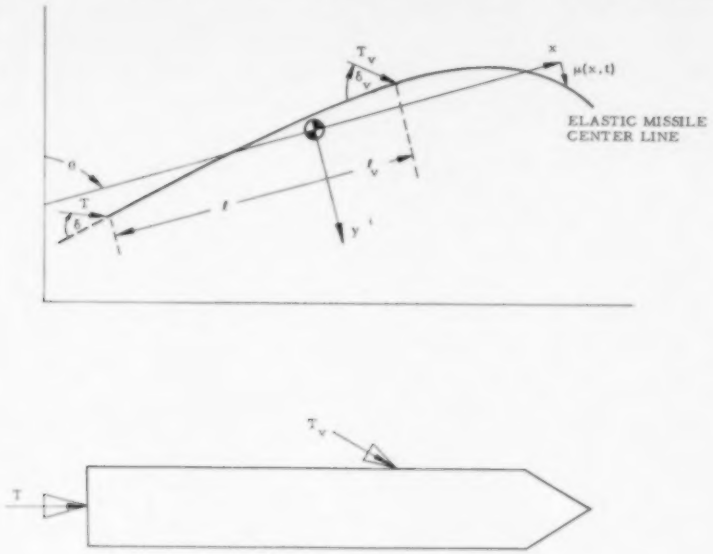


Fig. 4. Missile geometry.

missile angular deviation from an inertial reference, U_v is the angle a gyro located at station v along the missile senses due to the structural bending, swivelling engine and sloshing effects.

Neglecting sloshing, aerodynamics and coupling to the main engine from inertial reaction forces, the following approximate relations result (the computer solution does not neglect these effects):

$$\theta = \frac{T_c l_c C_c \delta_c}{Is^2} = \frac{\mu_{c_c} C_c \delta_c}{s^2} \tag{2.2}$$

$$q_i = \frac{(C_c T_c / M)\,\phi_i(v)\delta_c}{s^2 + \omega_i^2}. \tag{2.3}$$

Structural damping is neglected entirely in equation (2.3). Using expressions (2.2) and (2.3) in equation (2.1) and with

$$U_v' = \sum_i q_i \phi_i'(v),$$

gives

$$\frac{\delta_v}{\delta_c} = C_v\left[\frac{\mu_{c_v}}{s} + s\sum_i \left(\frac{T_v}{M}\right)\frac{\phi_i(v)\phi_i'(v)}{s^2 + \omega^2}\right]. \tag{2.4}$$

In the case that $C_v\mu_{c_v}$, $C_v\phi_i\phi_i'$ have the same sign (for all i), then the right-hand side of equation (2.4) is a Foster function. Hence the zeroes and poles of δ_v'/δ_v must be complex conjugate pairs lying on the $j\omega$ axis and exhibit the separation property that the zeroes and poles must alternate. These statements are proved by letting $s = j\omega$ in equation (2.4):

$$-\frac{\delta_v'}{\delta_c}(j\omega) = j\left(\sum_{k,\,k=0} \frac{A_k\omega}{\omega_k^2 - \omega^2} - \frac{A_0}{\omega}\right) \tag{2.5}$$

where the A_k, A_0 are *positive* real numbers. The minus sign on the left-hand side of equation (2.5) is due to the assumption of positive feedback in the auxiliary loop. Then

$$\frac{d[-(\delta_v'/\delta_c)(\omega)]}{d\omega} = \sum_k \frac{A_k(\omega_k^2 + \omega^2)}{(\omega_k^2 - \omega^2)^2} + \frac{A_0}{\omega^2}. \tag{2.6}$$

Thus, the slope of $(-\delta_v'/\delta_v)$ as a function of real frequency is always positive. Hence every pole must be followed by a zero—the separation property. The zeroes must lie on the real frequency axis in conjugate pairs and from equation (2.4) one sees that the number of zeroes is one less than the number of poles.

The property that $\phi_i(X_1)\phi_i'(X_1)$, μ_{c_v}, have the same sign has been found to be true for almost every missile studied when X_1 is the nose or tail station. While suitable restrictive conditions have not yet been discovered that will allow a

rigorous proof, physical reasoning lends credibility to the specific cases verified. For if a force were applied normal to the missile center line at the nose or tail, the elastic compliance in the structure would lead one to expect that the total local rotation due to bending would have the same sense as the rotation of the whole body considered to be rigid. This property is *not* true in general for other arbitrary locations when all modes and all times of flight are considered (see Table 5). It may be possible, however, that other positions along the missile can station location if the rate gyro were located at the tail) is now apparent. More than 90° additional phase shift at any bending mode frequency could produce instability at the bending mode or give a negative damping factor in the mode. Also, the use of a large engine would result in an additional conjugate zero pair on the real frequency axis ("tail-wags-dog" zero), and this *could* destabilize the higher modes. As is shown below, only a small thrust, 1000 lb, is required to provide considerable mode damping. Hence, it appears within the realm of practicality to require

Table 2. Open Auxiliary Loop Poles and Zeroes, Nozzle and Rate Gyro at Tail
(First stage burnout, $T_v = 1000$, $C_v = 100$)

Dominant physical effect	Poles	Zeroes
Normal force	0·02	−0·03
Rigid body	0·14	
	−0·18	
First bending mode	$-0{\cdot}23 \pm j25{\cdot}2$	$-0{\cdot}10 \pm j17{\cdot}1$
Second bending mode	$-1{\cdot}18 \pm j64{\cdot}3$	$-0{\cdot}33 \pm j45{\cdot}9$
Third bending mode	$-2{\cdot}47 \pm j88{\cdot}2$	$-0{\cdot}64 \pm j84{\cdot}9$
First stage hydraulics	−56·5	−58·7
	$-9{\cdot}9 \pm j91{\cdot}1$	$-10{\cdot}8 \pm j82{\cdot}5$
Second stage hydraulics	$-11{\cdot}1 \pm j36{\cdot}5$	$-11{\cdot}2 \pm j36{\cdot}4$
	$-38{\cdot}2 \pm j77{\cdot}9$	$-38{\cdot}3 \pm j77{\cdot}7$
Rate gyro	$-75{\cdot}1 \pm j130$	$-75{\cdot}1 \pm j130$
Position gyro	−200	−200
Mag. amp.	−308	−308
Filter	−9·0	−9·0

be found in specific instances, particularly when only three or so modes need be considered.

Now that the poles and zeroes of δ_v'/δ_v are known to lie on the real frequency axis and alternate and a simple pole is at the origin, it follows at once that the auxiliary loop is stable for all values of gain C_v as long as no additional dynamics are present in the loop. This result is subject to the terms of the approximations of equations (2.2) and (2.3). However, computer results for the more complete system representation show the assumptions and conclusions to be appropriate. (See Table 2 and Fig. 5, the open auxiliary loop Bode diagram.)

The reason for the use of an auxiliary thrust device instead of the main engine (the use of the main engine would satisfy the requirement of a nozzle that has dynamic effects only at frequencies higher than the highest bending frequency of interest. For instance, one might control the amount of gas flow out of nozzles placed normal to the missile center line instead of swivelling the nozzles. Further justification of this thesis is beyond the intent of this paper.

Actually some additional lag is desirable to help cut off the auxiliary loop. This lag may be the rate gyro natural dynamics, engine dynamics, filters or shockmount dynamics. The Bode diagram for δ_v'/δ_v at the burnout condition is shown on Fig. 5. A double lag of 100 r.p.s. has been added to produce greater attenuations at the higher frequencies.

It may very well turn out that the main engine could be used instead of the verniers in some

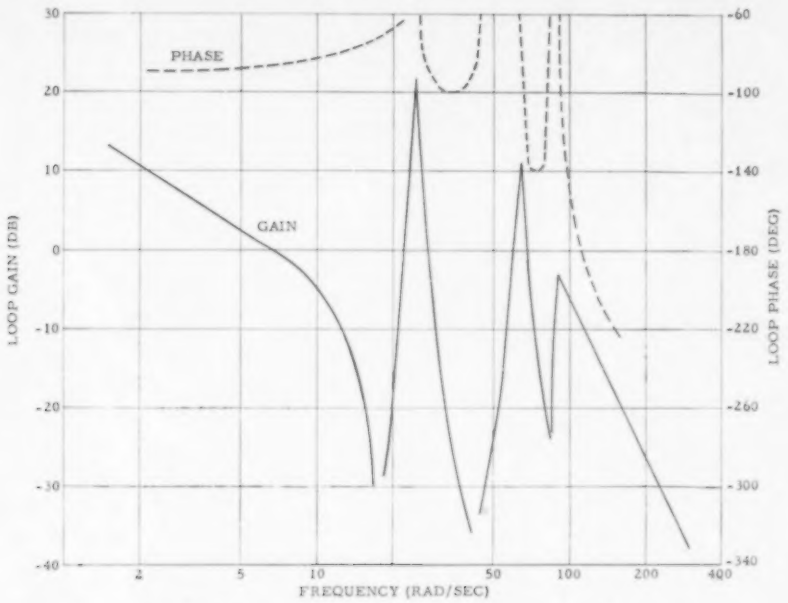


Fig. 5. Open auxiliary loop Bode plot, burnout condition $C_v = 100\ \text{sec}^{-1}$, double lag $[1/1 + (s/100)]^2$.

instances. However, if this is the case, it probably would not be necessary to have a second rate gyro since the stabilization problem of the auxiliary loop is now identical to the original stabilization problem. In fact, the only new idea introduced in this paper is the use of an auxiliary thrust system whose dynamics permit a loop to be closed around an elastic missile with assured stability margin.

It has been demonstrated that the auxiliary loop is always stable under the conditions described above. It will now be demonstrated that the zeroes of $(\delta_v'/\delta_v) - 1 = 0$ are indeed the poles of δ_t/δ_c the "main loop" transfer function. Denote the "main loop" transfer function by $\delta_t/\delta_c = |A(s)|/|B(s)|$ with $\delta_v \equiv \delta_v'$. Here $|A(s)|$ and $|B(s)|$ are determinants whose elements are polynomials in s. Note that $|B(s)|$ is a function only of the system and does not depend on δ_c. Similarly, the auxiliary loop transfer function is denoted by

$$\frac{\delta_v'}{\delta_v} = \frac{|C(s)|}{|D(s)|} \quad \text{with} \quad \delta_c = 0.$$

Imposing the conditions $\delta_c = 0$ and $\delta_v' = \delta_v$ on both representations, we determine values of s such that

$$|B(s)| = 0, \tag{2.7}$$

and values of s such that,

$$|D(s)| - |C(s)| = 0. \tag{2.8}$$

Since both representations now determine poles of the same system, this implies that at least some (actually all) of the "main loop" poles, values of s that satisfy equation (2.7), can be obtained from the auxiliary loop system as solutions of the equation $|D(s)| - |C(s)| = 0$, or equivalently, $(\delta_v'/\delta_v) - 1 = 0$. In fact, it can be shown from the governing system of equations that the determinants B, C and D differ only in elements of one column and that the latter two combine to give the exact relation

$$|D(s)| - |C(s)| \equiv |B(s)|. \tag{2.9}$$

The result, that the poles which would seem to appear only in the main loop (such as filter poles) still appear in the auxiliary loop poles, is a consquence of the fact that the corresponding

terms were not factored out in forming the auxiliary loop equations. Such poles are covered by zeroes in the auxiliary loop transfer function.

Since the poles of δ_v'/δ_v include the bending poles on the $j\omega$ axis (neglecting the small damping in the structure and that coupled from the hydraulics) and since $(\delta_v'/\delta_v) - 1 = 0$ has roots only in the left half plane, then the use of the auxiliary feedback always results in improved bending mode damping. See Tables 1, 3 and 4 for a comparison of bending damping with and without the auxiliary feedback. In all cases computed, a vernier thrust of 1000 lb was assumed.

3. Methods for Obtaining Required Damping

From the δ_v'/δ_v transfer function, an estimate of the gain C_vT_v required for a given closed loop damping in a particular bending mode can be obtained. For instance, the maximum damping in the third bending mode without filtering is approximately

$$\zeta = \frac{\omega_{p_3} - \omega_{z_3}}{\omega_{p_3} + \omega_{z_3}} \tag{3.1}$$

where ω_{p_3} is the third bending pole frequency, ω_{z_3} is the δ_v'/δ_v zero below ω_{p_3}. This relation derives from the fact that the locus branch connecting the pole–zero pair is an arc of a circle if the phase shift β from the rest of the system poles and zeroes is constant in the region of the pole–zero pair. The center of this circle is at a point in the S plane given by

$$j\omega_0 = \frac{j(\omega_{p_3} + \omega_{z_3})}{2} \tag{3.2}$$

$$\sigma_0 = -\frac{(\omega_{p_3} - \omega_{z_3})}{2 \tan \beta} \tag{3.3}$$

where β has positive phase contributions from zeroes and negative phase contributions from poles. The radius of the circle is

$$R_0 = \frac{(\omega_{p_3} - \omega_{z_3})}{2} \operatorname{cosec} \beta. \tag{3.4}$$

From Fig. 6, an illustrative root locus plot of equation (2.4), which is the approximate open auxiliary loop transfer function, it is seen that some additional lag will actually make possible a greater damping than given by equation (3.1). The lag cutoff filter should not come in before the Bode diagram of Fig. 5 indicates the loop gain is well attenuated; otherwise, some higher

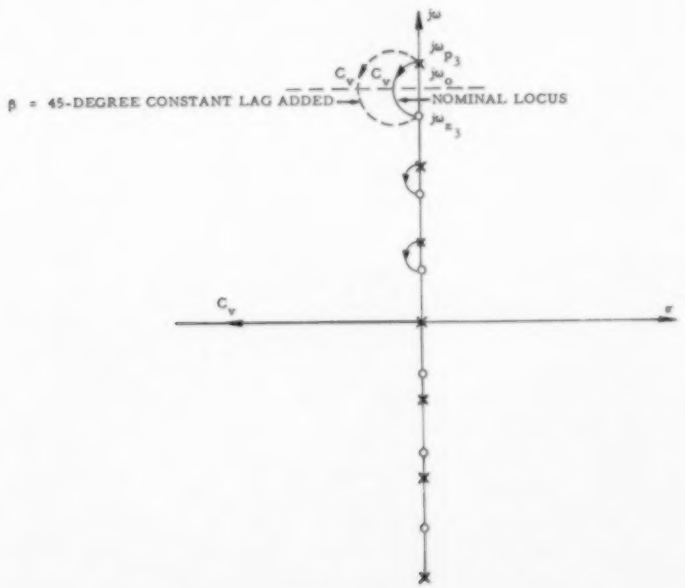


Fig. 6. Schematic root locus, open auxiliary loop.

mode not accounted for in the analysis may not be either amplitude or phase stabilized. The lag can be chosen to give a maximum damping for a given loop gain C_v. Tables 3 and 4 present the closed auxiliary loop poles as a function of location, C_v, and corner frequency of lag networks. With $T_v = 1000$, $C_v = 100$ appears to be a reasonable compromise between the closed loop damping obtained and stability margins of the auxiliary loop. However, no simple method is available that will permit independent specification of each mode damping, loop gain, and still retain simple cutoff filtering.

Table 3. Maximum Aerodynamic Pressure Condition

Location	C_v	Filter	Mode	Location of poles in S plane	Damping in per cent of critical	Damping in per cent of critical due to aux. loop
Tail	100	—	1	$-3{\cdot}7 \pm j18{\cdot}0$	20·2	19·0
			2	$-1{\cdot}5 \pm j35{\cdot}9$	4·2	3·2
			3	$-1{\cdot}5 \pm j62{\cdot}3$	2·4	1·3
Tail	100	$\omega_1 = \omega_2 = 100$ r.p.s.	1	$-3{\cdot}9 \pm j19{\cdot}7$	19·4	18·2
			2	$-1{\cdot}5 \pm j37{\cdot}1$	4·1	3·1
			3	$-1{\cdot}4 \pm j62{\cdot}7$	2·2	1·1
Tail	150	$\omega_1 = \omega_2 = 100$ r.p.s.	1	$-6{\cdot}8 \pm j20{\cdot}7$	31·5	30·3
			2	$-2{\cdot}6 \pm j37{\cdot}4$	7·0	6·0
			3	$-1{\cdot}5 \pm j62{\cdot}8$	2·4	1·3
Tail	200	$\omega_1 = \omega_2 = 100$ r.p.s.	1	$-10{\cdot}8 \pm j24{\cdot}4$	40·5	39·3
			2	$-4{\cdot}1 \pm j37{\cdot}5$	10·9	9·9
			3	$-1{\cdot}6 \pm j62{\cdot}9$	2·5	1·4
Nose	−100	—	1	$-6{\cdot}7 \pm j21{\cdot}6$	30·0	28·8
			2	$-10{\cdot}4 \pm j36{\cdot}2$	27·6	26·6
			3	$-2{\cdot}4 \pm j61{\cdot}9$	3·9	2·8

Comparisons of the Bode diagrams for δ_v/δ_c with and without feedback (Figs. 2 and 7, 3 and 8) emphasize the significant decrease in peaking at the bending resonant frequencies. Indeed, this is the major effect sought by this method. While the amount of damping injected into the bending modes decreases from the first to third mode, this is compensated by the fact that the system can be more easily attenuated at the higher modes. No simple explanation is offered for the apparent preference of the damping for the lower modes, but the system tends to be self-regulatory in the sense that more damping is obtained when the bending is more pronounced (and when bending is more of a problem).

4. Low Frequency Effects of Auxiliary Loop

The value of auxiliary loop gain influences not only the damping in the bending modes, but also the rigid body dynamics of the missile in a beneficial way. Considering the missile to be a rigid body, with the vernier nozzle angle

$$\delta_v = -C_v s\theta, \quad \text{then}$$
$$Is^2\theta = (Tl)\delta - T_v l_v C_v s\theta. \tag{4.1}$$

Note that the product $l_v C_v$ is always made positive by an appropriate choice of the sign of C_v. The rigid body transfer function from equation (4.1) is then

$$\frac{\theta}{\delta} = \frac{-(Tl)/I}{s[s + (T_v l_v C_v)/I]}. \tag{4.2}$$

Table 4. Burnout Condition

Location	C_v	Filter	Mode	Location of poles in S plane	Damping in per cent of critical	Damping in per cent of critical due to aux. loop
Tail	10	—	1	$-0.55 \pm j25.2$	2.2	1.3
			2	$-1.8 \pm j64.2$	2.8	1.4
			3	$-2.5 \pm j87.9$	2.8	0.2
Tail	30	—	1	$-1.2 \pm j25.1$	4.8	3.9
			2	$-2.9 \pm j64.0$	4.5	3.1
			3	$-2.5 \pm j87.5$	2.8	0.2
Tail	50	—	1	$-1.8 \pm j25.0$	7.2	6.3
			2	$-4.0 \pm j63.8$	6.3	4.9
			3	$-2.4 \pm j87.1$	2.9	0.3
Tail	75	—	1	$-2.6 \pm j24.8$	10.5	9.6
			2	$-5.5 \pm j63.4$	8.7	7.3
			3	$-2.3 \pm j86.8$	2.7	0.1
Tail	100	—	1	$-3.5 \pm j24.4$	14.2	13.3
			2	$-6.9 \pm j63.0$	10.9	9.5
			3	$-2.3 \pm j86.6$	2.7	0.1
Tail	200	—	1	$-5.9 \pm j21.2$	26.8	25.9
			2	$-13.2 \pm j58.7$	22.8	20.6
			3	$-1.9 \pm j85.9$	2.2	−0.4
Tail	100	$\omega_1 = \omega_2 = 75$	1	$-3.1 \pm j27.2$	11.3	10.4
			2	$-2.0 \pm j67.7$	2.9	1.5
			3	$-4.2 \pm j87.9$	4.8	2.2
Tail	100	$\omega_1 = \omega_2 = 100$	1	$-3.6 \pm j26.6$	13.4	12.5
			2	$-3.3 \pm j68.1$	4.8	3.4
			3	$-4.0 \pm j87.2$	4.6	2.0
Tail	100	$\omega_1 = \omega_2 = 150$	1	$-3.9 \pm j25.8$	15.0	14.1
			2	$-5.3 \pm j67.8$	7.8	6.4
			3	$-3.5 \pm j86.6$	4.0	1.4
Nose	−100	—	1	$-21.1 \pm j30.5$	57.0	55.8
			2	$-28.5 \pm j40.0$	58.0	57.0
			3	$-5.3 \pm j73.1$	7.2	4.6

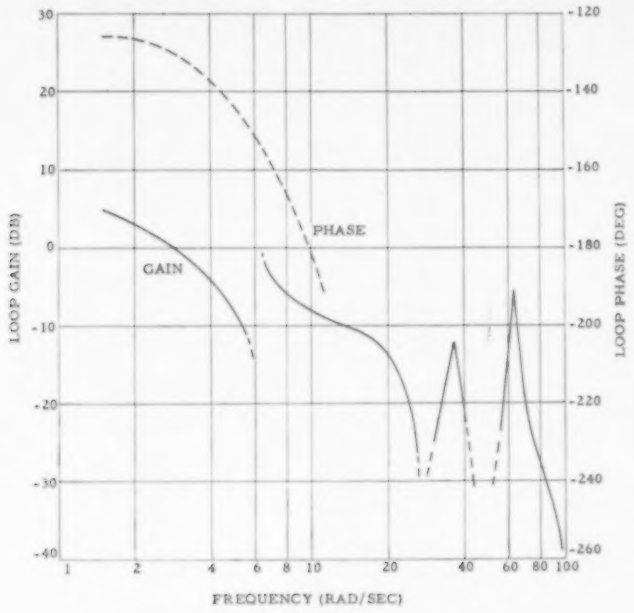


Fig. 7. Bode plot δ_i/δ_C, with auxiliary loop, $C_v = \dfrac{100\ \text{sec}^{-1}}{[1 + (s/100)]^2}$, maximum aerodynamic pressure condition.

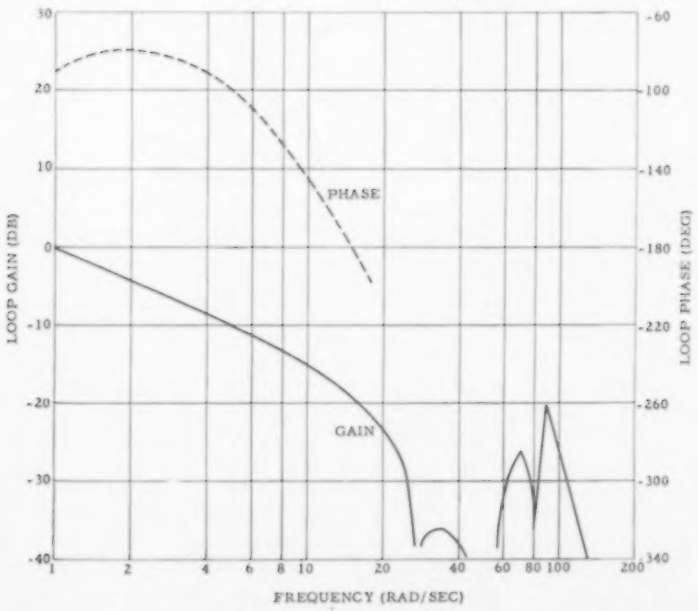


Fig. 8. Bode plot δ / $_c$ with auxiliary loop at tail $C_v = 100$, burnout condition.

Since this transfer function is stable for all values of gain, the need for the original rate gyro is certainly decreased. Even with the complete system representation this is true as comparison of the phase margins with and without the auxiliary feedback reveals (see Figs. 3 and 7). Less rate gyro signal in the "main loop" would decrease the amplitude of the high frequency peaks and thus give the same result as increasing the bending damping.

It should be noted that the gain constant in equation (4.2) is the same value obtained without auxiliary feedback, $T_v = 0$. As will be shown, this is very nearly true in general. The main loop transfer function can be written in the form

$$F(s) = \frac{|A(s)|}{|B(s)|} = a\pi_i(s - z_i)/b\pi_j(s - p_j) \quad (4.3)$$

where the z_i and p_j are the open loop poles and zeroes. The open loop Bode gain is obtained by evaluating this function for $s = 0$ (poles and zeroes at the origin being first removed):

$$K_B = \frac{|A(0)|}{|B(0)|} = a\pi(-z_i)/(b\pi(-p_j)). \quad (4.4)$$

In this case, equation (2.1) shows that the angle δ_v' is zero, and the thrust nozzle is acting only through the deflection slope at the nozzle location. Now the change in the Bode gain due to the addition of the nozzle is of the order T_v/T, where T_v is the thrust of the nozzle and T is the thrust of the main engine. Since this ratio will be very small in general (approximately 0·003 for cases now under consideration), the Bode gain is for practical purposes essentially unaffected by the addition of the thrust nozzle. This independence of the Bode gain constant to the presence or absence of the auxiliary loop is dependent on the condition that $s = 0$ is not a zero of $A(s)$ or $B(s)$. While this condition is not satisfied for the simplified example, (4·2), it is true for the actual system.

Furthermore, the coefficients a and b are determined by the terms of highest degree in s in the mathematical systems. Since the auxiliary loop is of lower degree in s than the main loop, the root-locus gain a/b is hence unaffected by the addition of the nozzle. If filters introduced into the auxiliary system are represented as polynomials in s multiplying δ_v' in equation (2.1), then a and b are altered in the same proportion so that their ratio is still unaltered. The preceding results then imply that the ratio of the product of the zeroes to the product of the poles also remains essentially independent of the nozzle addition.

5. Linear Range of Auxiliary Loop

Aside from stability considerations, a low C_v is desirable so that the maximum nozzle travel of 90° (more travel would actually decrease the damping force) corresponds to some reasonable missile angular velocity. If 90° is to be 10 times the rate gyro threshold, then

$$C_v = \frac{90 \text{ deg}}{10 \times \text{rate gyro threshold}}. \quad (5.1)$$

The nozzle thrust required is the gain required to produce the desired closed loop damping divided by C_v. Saturation of the auxiliary loop would have the effect of decreasing C_v from the linear range value; nonetheless, the damping will still be greater than the original system, $C_v = 0$. With $C_v = 100$, the damping system would saturate at a missile rate of 0·9 deg/sec. If the thrust were doubled and the loop gain held constant, the saturation level would double.

6. Practical Considerations

The practical problem of locating a rate gyro in the vibration environment of the engine compartment, or what is probably less practical, locating even small nozzles at the nose, will not be minimized. It should be noted that the bandwidth requirements placed on the auxiliary loop coupled with the inherent necessity for a small thrust device so that this high bandwidth can be realized make this loop sensitive to noise from the rate gyro.

The picture is not entirely pessimistic, however, since some filtering in the auxiliary loop is actually desirable. A shockmount for the rate gyro might be used or advantage taken of the natural cutoff of rate gyros. In this study a double lag at 16 c/s was found to yield good results for a system stability point of view.

Table 5. Non End Point Stations

Time	Location station number, inches from nose	C_v	Filter	Mode	Pole location in S plane	Damping
Maximum aerodynamic pressure condition	141	−100	—	1	$-3{\cdot}7 \pm j18{\cdot}9$	19·2
				2	$-4{\cdot}9 \pm j34{\cdot}5$	14·0
				3	$-1{\cdot}6 \pm j62{\cdot}3$	2·6
Maximum aerodynamic pressure condition	432	−100	—	1	$+0{\cdot}48 \pm j18{\cdot}6$	−2·6
				2	$-1{\cdot}1 \pm j36{\cdot}3$	3·0
				3	$-0{\cdot}82 \pm j62{\cdot}7$	1·3
Burnout condition	141	−100	—	1	$-6{\cdot}3 \pm j24{\cdot}0$	25·4
				2	$-3{\cdot}1 \pm j63{\cdot}6$	4·9
				3	$+0{\cdot}3 \pm j89{\cdot}3$	−0·34

Comparison of Tables 3 and 4 show the nose location to give more damping for a given C_v, T_v. This follows from the fact that the missile bends more at the nose in general. However, in practice, the choice of nose or tail location will probably be dictated by weight considerations.

III. CONCLUSIONS

The principal results indicated in the preceding discussion to be obtained by the addition of a small thrust device and rate gyro may be summarized as follows:

(1) Damping in the bending modes can be increased in general by the use of an auxiliary feedback loop consisting of a rate gyro and thrust device, whose thrust is small in comparison with the main engine thrust, provided that

(a) The rate gyro and thrust device are located together at either missile nose or tail;

(b) the thrust device has no dynamic effects below the highest bending frequency of importance. This frequency is determined by the numerical values of the system parameters.

Specific cases are not precluded where (a) and/or (b) are unnecessary.

(2) Under conditions stated, the auxiliary loop is generally stable.

(3) Application of this method to a representative of I.C.B.M. autopilot design eliminates missile structural bending as a limiting factor in the stability margins obtainable in this autopilot design.

APPENDIX

Basic equations governing the planar motion of a missile are derived in Ref. (4), and equations describing the missile control system are given in Ref. (5); consequently, they will not be repeated here. Principally, Lagrange's equations of motion are derived using generalized coordinates: θ and $\bar{a}_y$ representing the missile attitude and lateral acceleration, q_1, q_2 and q_3 representing the first three natural bending modes of the missile, ζ_1 and ζ_2 representing the fuel sloshing in each of two tanks, and δ representing the angle the first stage engine makes with the missile axis.

In this investigation these equations were modified to include the dynamics of the second stage engine with its associated hydraulic system and the additional thrust nozzle with its associated rate gyro. The inclusion of the second stage engine requires that the deflection $u(x, t)$ of the missile in bending given by

equation (4.9) of Ref. (4) be modified to include the additional term

$$(K_{02} - K_{12}x)\delta_2$$

where

$$K_{02} = m_{c2}\bar{z}_2/M$$

$$K_{12} = \frac{1}{I}(I_{h2} + m_{c2}\bar{z}_2 l_{02}).$$

Here M and I are the mass of the missile and its moment of inertia about the center of gravity, m_{c2} and I_{h2} are the mass of the second stage engine and its moment of inertia about its pivot, $\bar{z}_2$ is the distance between the center of gravity and the pivot point of the second engine, l_{02} is the distance between the engine pivot and the center of gravity of the missile (considered positive if the engine is aft of the center of gravity) and δ_2 is the angle that the engine makes with the missile axis. Similarly, equation (6.5) of Ref. (4) giving the total kinetic energy of the missile is modified to include the additional term

$$T^*_{\delta 2} = \tfrac{1}{2}\int_0^{b_2}\{z^2\dot{\delta}_2^2 - 2z\dot{\delta}_2(\bar{v}_y + x\dot{\theta} + \dot{u}) + z^2\delta_2^2\dot{\theta}^2 + 2z\delta_2\theta(\bar{v}_x - u\dot{\theta})\}\,dm_{c2},$$

representing the additional kinetic energy of the second stage engine due to its swivelling through the angle δ_2. The integration is performed over the length b_2 of the engine. Lagrange's equations of motion of Ref. (4) thus contain additional terms involving the variable δ_2 arising from the kinetic energy $T^*_{\delta 2}$ and the redefinition of the deflection u. There is also the additional equation of motion for the second stage engine:

$$-m_{c2}\bar{z}_2(\bar{a}_y + \bar{v}_x\dot{\theta}) + (I_{h2} + m_{c2}\bar{z}_2 l_{02})\ddot{\theta} -$$

$$-(\ddot{\delta} - \delta\dot{\theta}^2)[m_{c2}\bar{z}_2 K_0 + K_1(I_{h2} + m_{c2}\bar{z}_2 l_{02})] +$$

$$+(\ddot{\delta}_2 - \delta_2\dot{\theta}^2)\{I_{h2} - m_{c2}\bar{z}_2 K_{02} - K_{12}(I_{h2} +$$

$$+ m_{c2}\bar{z}_2 l_{02})\} + \sum_{j=1}^{2}(\zeta_j - \zeta_j\dot{\theta}^2)[m_{c2}\bar{z}_2 K_{2j} -$$

$$- K_{3j}(I_{h2} + m_{c2}\bar{z}_2 l_{02})] - \sum_{i=1}^{3}$$

$$(\ddot{q}_i - q_i\dot{\theta}^2)\left[\int_0^{b_2} z\phi_i\,dm_{c2} + K_{12}I\sum_{j=1}^{2}K_{4j}\phi_i'(l_j)\right] +$$

$$+ m_{c2}\bar{z}_2(\bar{a}_x + g\cos\theta)[\delta_2 + u'(l_{02})] = Q_{\delta 2}$$

with

$$Q_{\delta 2} = r_2 K_{s2}(\delta_{c2} - \delta_2) - D_{m2}\delta_2 - K_{f2}\delta_2$$

where

r_2, K_{s2}, δ_{c2}, D_{m2}, K_{f2} are as defined in Ref. (5) only for the second stage engine hydraulic system. In a manner exactly analogous to that for the first stage system, the additional control equations for the second stage engine and hydraulic system are taken as follows:

$$m_{r2}\ddot{x}_{r2} + d_{r2}\dot{x}_{r2} + c_{r2}x_{r2} = K_{r2}(\bar{K}_{32}\ddot{e}_2 + \bar{K}_{31}\dot{e}_2 + e_2)$$

$$A_{H2}r_2\dot{\delta}_{c2} = K_{r2}x_{r2} - \frac{L_2 K_{s2}}{A_{H2}}(\delta_{c2} - \delta_2) -$$

$$-\frac{V_{H2}}{B_2}\frac{K_{s2}}{A_{H2}}(\dot{\delta}_{c2} - \dot{\delta}_2) = 0.$$

$$e_2 + \delta_{c2} = 0.$$

In this formulation no command signal is being supplied to the second stage engine hydraulic system, and the system is operative only through the inertia of the engine. As a consequence, it is emphasized that the inclusion of the second stage engine and hydraulic system has only a very minor influence on the results of such an investigation and could very well be neglected without altering the conclusions.

The inclusion of the additional thrust nozzle —the principal purpose of this study—is achieved by assuming that the nozzle is at the position $x = x_{ve}$ on the missile and provides a thrust T_v acting at a variable angle δ_v with the missile axis. Assuming this nozzle to be relatively small, its inertia is neglected; consequently, the only effect of this additional thrust is to add additional generalized forces in Lagrange's equations of motion. These additional generalized forces are as indicated:

Equation of lateral motion:

$$T_v[\delta_v + u'(x_{ve})]$$

Moment equation about the center of gravity:

$$T_v\{[\delta_v + u'(x_{ve})]\,x_{ve} - u(x_{ve})\}$$

Bending equations:

$$T_v[\delta_v + u'(x_{ve})]\left(\frac{\partial u}{\partial q_i}\right)_{x=x_{ve}} \quad i = 1, 2, 3$$

The angle δ_v of the nozzle is assumed to be controlled by a rate gyro, mounted at the same position as the thrust nozzle, which senses the rate of rotation of the missile at that location. In terms of Laplace transform variables and transform parameter s, this control is expressed by the following relation:

$$\delta_v = c_v \frac{s[\theta + u'(x_{v_e})]}{[1 + s/(\omega_{v1})][1 + (s/\omega_{v2})]}.$$

The denominator on the right-hand side is included to represent possible filtering in the auxiliary system or to represent possible dynamic effects of a shockmount for the gyro.

The total system is represented by twenty-four non-linear ordinary differential equations of at most second order in the variables representing the missile dynamics, autopilot and auxiliary thrust systems. For the purposes of the stability analyses of the present investigation, the missile's motion is assumed perturbed about a given nominal trajectory. The governing equations are then linearized with respect to the variables representing the perturbed motion. However, since the resulting equations have variable coefficients, the stability of the missile and autopilot system is investigated at two times of flight: 60 sec after launching at the time of maximum aerodynamic pressure and 130 sec after launching at the time of first stage burnout. The coefficients are then assumed to have the constant values they have at these times. Finally, the Laplace transform of each equation is taken, assuming zero initial conditions.

Because of the magnitude of the resultant mathematical problem, a complete analytical solution of the equations is virtually impossible. Thus numerical solutions for the poles and zeroes of the autopilot open loop and auxiliary open loop transfer functions are obtained using the Generalized Eigenvalue Problem Package RWGE21 of the SHARE library programmed for the IBM 704 digital computer.

Glossary of Terms

T Thrust of main engine.

T_v Thrust of auxiliary engine.

l Distance from main engine attachment point to missile c.g.

l_v Distance from auxiliary engine attachment point to missile c.g.

I Moment of inertia of rigid missile about its center of gravity.

C_v Angular displacement of auxiliary engine from neutral per unit missile angular rate at the auxiliary engine station.

θ Angular displacement of rigid missile center line from an inertial reference.

U_v Bending displacement of station $x = v$.

U'_v Local slope due to bending displacement at station $x = v$.

$\phi_i(x)$ i^{th} bending mode eigenfunction.

$\phi'_i(x)$ $d\phi_i(x)/dx$.

q_i Amplitude factor of i^{th} eigenfunction.

u_{v_e} Control coefficient for auxiliary loop, $T_v l_v / I$

M Missile mass.

ω_i i^{th} bending mode frequency.

REFERENCES

1. W. A. Dite, L. E. Freed and L. F. Lanzer, *WS107A-2, Series I, Lot B, Stage I Autopilot Design Study*, Space Techn. Lab. Rep. GM-TN-0165-00139 (May 1958).
2. R. van Valkenberg, *The Difficulty of Obtaining a Shaping Network to Approximate Desired Magnitude and Phase Characteristic for Missile Design*, Space Techn. Lab. Rep. GM-TN-0165-00304 (Oct. 1958).
3. G. A. Becky, D. R. Breuer, L. E. Freed and H. C. Thomson, *An Investigation of Non-Linear Autopilot Filters for the Titan Missile*, Space Techn. Lab. Rep. GM-TM-59-0000-00037 (Feb. 1959).
4. D. Young, *Generalized Missile Dynamic Analysis II —Equations of Motion*, Space Techn. Lab. Rep. GM-TR-0165-00359 (Apr. 1958).
5. D. C. Martin, *Generalized Missile Dynamic Analysis VI—Control System Equations*, Space Techn. Lab. Rep. GM-TR-0165-000359 (Apr. 1958).

Acknowledgement—The authors wish to thank Mr. Raymond F. Kramer and Mr. Robert B. Freund of the Mathematical Analysis Department, Computation and Data Reduction Center, Space Technology Laboratories, Inc., for their helpful comments on various aspects of this investigation and for their extremely valuable assistance in obtaining the numerous numerical computations required.

A NAVIGATION THEORY FOR ROUND-TRIP RECONNAISSANCE MISSIONS TO VENUS AND MARS

RICHARD H. BATTIN and J. HALCOMBE LANING, JR.
Instrumentation Laboratory, Massachusetts Institute of Technology

Abstract—A self-contained navigation scheme is described and analysed for the case of an unmanned spaceship launched from Earth and established in a free-fall solar orbit that passes within a few thousand miles of another planet and subsequently returns to Earth. A statistical study of the navigation errors and the micro-rocket fuel requirements is made, using a three-dimensional model of the solar system. Several different trajectories are subjected to a systematic study to determine the optimum times at which positional checks and velocity corrections should be made in order to minimize both the amount of fuel required and the errors in passing distance at the target planet and miss distance upon return to Earth.

1. INTRODUCTION

The technical feasibility of an unmanned photographic reconnaissance flight to the planet Mars was demonstrated in a recent paper by Laning *et al.*[1] The mission which they consider consists of launching a vehicle from Earth into a predetermined free-fall solar orbit. The vehicle makes brief contact with Mars, passing within a few thousand miles of the Martian surface, and, ultimately, returns to Earth with scientific data packaged to survive the shock of atmospheric re-entry. It is the purpose of the present paper to explore in detail the navigational aspects of such a venture. Indeed, the concepts which we shall develop are fairly general and may be adapted to the interplanetary guidance problem for a variety of space missions.

Unfortunately, until we have at our disposal more powerful sources of energy than are now available in the form of chemical fuels, space vehicles must be confined to travel the admittedly restrictive class of free-fall paths. In the navigation theory presented here it is assumed that the vehicle is acted upon only by gravitational forces during its voyage, with the sole exception of the brief periods of acceleration associated with navigation velocity corrections. Thus, the theory is not directly applicable to vehicles undergoing a continuous low thrust throughout the entire trip.

The basis for the mission is the selection of a suitable free-fall path for the spaceship to follow. In order to assure success, the vehicle must not be permitted to deviate substantially from the planned course because of the limited reserve of energy which can be transported. Again, because of fuel requirements, we consider only nonstop, round-trip trajectories. Although the idea of orbiting a planet for an extended period is appealing, it is felt that much useful data could be acquired by a vehicle during a single pass of the destination planet.

It is possible to establish a round-trip course to the planets Venus or Mars with a departure velocity from Earth which is only slightly larger than the minimum escape velocity. The mathematical details, together with the complete computational procedure needed to determine such interplanetary trajectories, have been given in a recent paper[2] by one of the present authors and will not be considered further here.

In our navigation theory we do not deal with the problem of launch guidance from Earth. It is assumed that the main propulsion stages are completed at the start of the problem. In all other respects, however, we treat the guidance problem from the time the vehicle leaves Earth until it returns.

In order to minimize the requirements of the navigation system, no attempt is made to solve the trajectory problem during flight. Instead, the method of navigation selected is based on a perturbation theory in which only positional deviations from a reference path at certain

predetermined times are utilized. Specifically, the navigation problem is solved on board the spaceship by means of (1) a sequence of optical measurements of angles between lines of sight to various celestial objects; (2) a clock; (3) a digital computer to determine positional deviations from the celestial observations and to calculate both the required velocity corrections and a correction to be applied to the computer clock; and (4) a micro-rocket propulsion system for making small changes in the spaceship velocity as directed by the computer.

For navigation on both the outbound and return portions of the trip, a number of check-points is assumed at which positional deviations are determined from celestial observations of the type described in Section 2. It is assumed for the present that a velocity impulse of any magnitude and direction may be achieved at any of the check-points. Later in Section 5, attention is given to the selection of check-points to determine, in a statistical sense, the effect on fuel consumption and guidance accuracy of the following sources of error: (1) initial velocity errors after termination of the main propulsion, attributable to launch guidance; (2) optical sighting errors; (3) errors in establishing desired velocity corrections; and (4) drift errors in the clock. The results of this statistical study, as applied to round-trip reconnaissance missions to both Venus and Mars, are presented in Section 5. Throughout the analysis a three-dimensional model of the solar system is used.

2. THE NAVIGATIONAL FIX

As the first step in the formulation of a guidance theory, we consider the processes involved in determining spaceship position by means of a celestial fix. The physical operation involved in the fix comprises a sequence of measurements of the angles between selected pairs of celestial bodies, together with the measurement of the angular diameter of a nearby planet when appropriate. One further observation is implied in the fix; namely, the recording of time as indicated by the spaceship clock. The intended result of these observations is the determination of coordinates of spaceship position together with a correction to the current clock reading. Here we shall describe one possible form of the required calculations and then relate the resultant errors in position and time estimates to the errors in the primary measurements.

It is assumed throughout the analysis that approximations to spaceship position and to time are already known so that perturbation techniques may be employed. In the application considered in this paper, no real restriction is implied by this assumption since deviations from a selected reference trajectory must be kept small in order to complete the mission with a reasonable fuel supply. Specifically, we assume the existence of a reference time T for the fix and a reference position vector $\mathbf{r}$ for the spaceship at time T. We further assume exact knowledge of the position and velocity of all relevant astronomical objects at time T.

No attempt is made to include within this theory such minor corrections as may arise from the finite speed of light, the finite distances of stars, or the departures of the motion of the planets from true ellipses. Such corrections can be lumped together for a particular reference point on the trajectory as a modification to the reference values for the various angles to be measured at that point.

To begin the analysis, we assume that when the spaceship clock indicates the reference time T, the sighting process commences. Let the clock be in error at this time by the amount δT so that the sighting begins at the true time $T - \delta T$. Now, since the sighting process consumes a non-negligible amount of time, we shall associate with each separate angle measurement a time increment δt representing the elapsed time from the start of the sighting process to the time at which that particular angle is measured. Therefore, the measurement is made at the true time $T - \delta T + \delta t$. We note that although the value of δt is accurately known to the spaceship computer, that of δT is not.

As the first type of measurement to be considered, we choose that of the angle from the Sun to an arbitrary planet. Subsequently, by allowing one or the other of these bodies to recede to an infinite distance, we may obtain

corresponding relations for the Sun–star or the star–planet type of measurement.

Let S_0 and P_0 be, respectively, the reference position of the spaceship and the position of a planet at time T. Let $\mathbf{r}$ be the vector from the Sun to S_0 and $\mathbf{z}$ the vector from S_0 to P_0. With A denoting the angle from the Sun line to the planet line, as shown in Fig. 1, we shall derive an expression for the change in angle δA arising from (1) the motion of the planet during the interval $\delta t - \delta T$ between the reference time and the actual time of measurement;

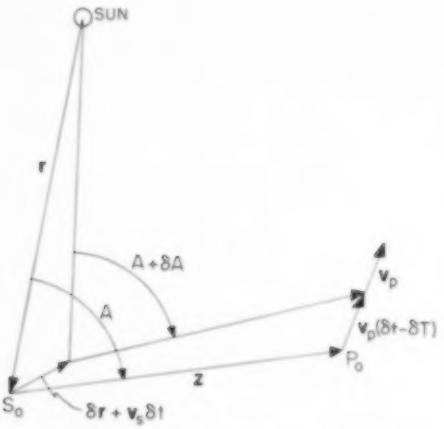


Fig. 1. Measurement of angles between celestial bodies.

(2) the initial displacement $\delta\mathbf{r}$ of the spaceship position with respect to S_0 at the time $T - \delta T$ when the sighting process begins; plus (3) the vector distance $\mathbf{v}_s\delta t$ travelled by the spaceship with velocity vector $\mathbf{v}_s$ from the start of the fix to the completion of the measurement of A. We note that the initial displacement $\delta\mathbf{r}$ may arise in part from the motion of the spaceship during the time interval δT and in part from its deviation from the reference position S_0 at time T. One of the significant tasks in analysing the fix data is properly to separate these components.

The angle A is given in terms of the scalar product

$$\cos A = -(\mathbf{r} \cdot \mathbf{z})/rz \tag{2.1}$$

where r and z denote the magnitudes of the respective vectors $\mathbf{r}$ and $\mathbf{z}$. Treating all changes as differentials, we may write

$$\delta(rz \cos A) = -(\mathbf{r} \cdot \delta\mathbf{z} + \mathbf{z} \cdot \delta\mathbf{r}') \tag{2.2}$$

where

$$\delta\mathbf{r}' = \delta\mathbf{r} + \mathbf{v}_s\delta t \qquad \delta\mathbf{z} = \mathbf{v}_p(\delta t - \delta T) - \delta\mathbf{r}' \tag{2.3}$$

and $\mathbf{v}_p$ is the velocity vector of the planet at time T. By expanding the left-hand side of equation (2.2) and noting that

$$\delta r = (\mathbf{r} \cdot \delta\mathbf{r}')/r \qquad \delta z = (\mathbf{z} \cdot \delta\mathbf{z})/z \tag{2.4}$$

we may obtain the following expression for δA:

$$\delta A = [\mathbf{n} - (\mathbf{n} \cdot \mathbf{m})\mathbf{m}] \cdot [\delta\mathbf{r}' - \mathbf{v}_p(\delta t - \delta T)]/z \sin A + \\ + [\mathbf{m} - (\mathbf{n} \cdot \mathbf{m})\mathbf{n}] \cdot \delta\mathbf{r}'/r \sin A \tag{2.5}$$

where we have denoted by $\mathbf{n}$ and $\mathbf{m}$, respectively, the unit vectors from S_0 toward the Sun and toward P_0.

By letting r or z become infinite in equation (2.5), we include all of the types of measurements made in this study except for the angular diameter of a nearby planet. Referring to Fig. 2, if D is the actual diameter of the planet, the apparent angular diameter A is found from

$$\sin (A/2) = D/2z \tag{2.6}$$

Taking differences as before, we find

$$\delta A = D\mathbf{m} \cdot [\delta\mathbf{r}' - \mathbf{v}_p(\delta t - \delta T)]/z^2 \cos (A/2) \tag{2.7}$$

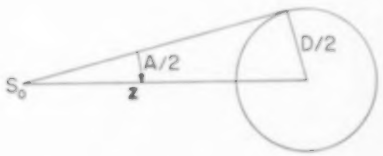


Fig. 2. Measurement of the apparent diameter of a planet.

The coefficients of δt in these expressions are numbers which may be computed in advance for each angle measured and stored in the spaceship computer. The resulting correction to δA may then be applied directly for each angle at the time the measurement is made. We assume this to be the case in our further work and drop the δt terms hereafter.

We shall assume that a total of $m (m \geq 4)$ measurements is made for a single fix. Then,

since the governing equations for each of the various types of angle measurements are linear in $\delta\mathbf{r}$, δT and δA, we may express the results in the matrix form

$$\delta\mathbf{A} = \mathscr{M}\mathbf{x} \tag{2.8}$$

where $\mathscr{M}$ is an $m \times 4$ matrix whose elements depend only upon the geometry of the configuration and upon the velocity vectors of the spaceship and planets. Also, $\delta\mathbf{A}$ is the m-dimensional column vector

$$\delta\mathbf{A} = \text{col}\,(\delta A_1, \delta A_2, \ldots, \delta A_m) \tag{2.9}$$

and $\mathbf{x}$ is the four-dimensional column vector

$$\mathbf{x} = \text{col}\,(\delta\mathbf{r}, \delta T) \tag{2.10}$$

For the remainder of the discussion it is necessary to distinguish between the measured or inferred values of the various quantities and their true values; e.g., $\delta\tilde{\mathbf{A}}$ will be the measured value of the deviation in $\mathbf{A}$ from its reference value and $\delta\mathbf{A}$ will be the true value of the deviation, etc.

When m is greater than four, some of the measurements are redundant. Therefore we may use a least-squares fit to the observed data to obtain a presumably more accurate estimate of position and time than would be possible with a minimum of four measurements. For this purpose, let σ_i and σ_t be, respectively, the estimated standard deviations of the ith angular measurement and the expected clock error at time T. Then, one approach to the determination of an estimate of the vector $\mathbf{x}$ is to find that vector $\tilde{\mathbf{x}}$ which minimizes the expression*

$$[\mathscr{S}(\mathscr{M}\tilde{\mathbf{x}} - \delta\tilde{\mathbf{A}})]^2 + (\delta\tilde{T}/\sigma_t)^2$$

where $\mathscr{S}$ is the mth order diagonal matrix whose elements are σ_1^{-1}, $\sigma_2^{-1}, \ldots, \sigma_m^{-1}$. The last term in this expression is included to account for the indirect measurement of time which results from the act of starting the fix when the spaceship clock indicates time T. The best direct measurement for $\delta\tilde{T}$ is, of course, zero.

* We remark that this least-squares procedure is equivalent to the "maximum likelihood" estimate if certain technical assumptions are made about the data.

Following the usual steps in the least-squares process, we are led to the following set of four simultaneous linear equations to be solved for the components of $\tilde{\mathbf{x}}$:

$$\mathscr{W}\tilde{\mathbf{x}} = \mathscr{M}'\mathscr{S}^2\delta\tilde{\mathbf{A}} \tag{2.11}$$

(A prime denotes the transpose of a matrix.) The fourth-order square matrix $\mathscr{W}$ is determined from

$$\mathscr{W} = \mathscr{M}'\mathscr{S}^2\mathscr{M} + \mathscr{Z} \tag{2.12}$$

where $\mathscr{Z}$ is the matrix whose only non-zero element σ_t^{-2} appears in the lower right corner. Thus, if we define

$$\mathscr{F} = \mathscr{W}^{-1}\mathscr{M}'\mathscr{S}^2 \tag{2.13}$$

then we have

$$\text{col}\,(\delta\tilde{\mathbf{r}}, \delta\tilde{T}) = \mathscr{F}\delta\tilde{\mathbf{A}} \tag{2.14}$$

We have already accounted for the fact that different individual observations occur at appreciable time intervals after the start of the fix; there remains to correct for the fact that the time of starting the fix is itself in error. The best estimate available to the spaceship computer indicates that the fix started $\delta\tilde{T}$ time units too soon. To obtain the deviations at the reference time T, indicated by $\delta\tilde{\mathbf{r}}(T)$, it is necessary to add to $\delta\tilde{\mathbf{r}}$ the vector distance travelled by the spaceship in time $\delta\tilde{T}$. If we define a 3×4 augmented matrix $\mathscr{X}$ as

$$\mathscr{X} = \|\mathscr{I}, \mathbf{v}_s\| \tag{2.15}$$

where $\mathscr{I}$ is the 3×3 identity matrix, then we have

$$\delta\tilde{\mathbf{r}}(T) = \mathscr{X}\mathscr{F}\delta\tilde{\mathbf{A}} \equiv \mathscr{Y}\delta\tilde{\mathbf{A}} \tag{2.16}$$

Finally, if we define an n-dimensional vector $\mathbf{f}$ to have components identical with the elements of the last row of the matrix $\mathscr{F}$, we have

$$\delta\tilde{T} = \mathbf{f}'\delta\tilde{\mathbf{A}} \tag{2.17}$$

To study the effects on navigation of the errors in the primary measurements, we write

$$\delta\tilde{\mathbf{A}} = \delta\mathbf{A} + \boldsymbol{\alpha}\ \delta\tilde{\mathbf{r}}(T) + \delta\mathbf{r}(T) + \boldsymbol{\epsilon}\ \delta\tilde{T} = \delta T + \tau \tag{2.18}$$

where $\delta\mathbf{A}$, $\delta\mathbf{r}$, δT are the true deviations and $\boldsymbol{\alpha}$, $\boldsymbol{\epsilon}$, τ are the errors made in the measurements

and the estimates. From equations (2.8), (2.12), (2.13) and (2.16), we have

$$\delta\tilde{\mathbf{r}}(T) = \mathscr{X}\mathscr{H}^{-1}\mathscr{H}'\mathscr{S}^{-2}(\mathscr{H}\mathbf{x} + \boldsymbol{\alpha})$$

$$= \mathscr{X}(\mathscr{S} - \mathscr{H}^{-1}\mathscr{Z})\mathbf{x} + \mathscr{Y}\boldsymbol{\alpha} \quad (2.19)$$

Therefore, if we define the four-dimensional vector $\mathbf{w}$ to have components identical with the elements of the last column of the $\mathscr{H}^{-1}$ matrix, equation (2.19) becomes

$$\delta\tilde{\mathbf{r}}(T) = \delta\mathbf{r}(T) + \mathscr{Y}\boldsymbol{\alpha} - \mathscr{X}\mathbf{w}\delta T/\sigma_t^2$$

Hence

$$\boldsymbol{\epsilon} = \mathscr{Y}\boldsymbol{\alpha} - \mathscr{X}\mathbf{w}\delta T/\sigma_t^2 \quad (2.20)$$

In a similar manner we may obtain

$$\tau = \mathbf{f}'\boldsymbol{\alpha} - w\delta T/\sigma_t^2 \quad (2.21)$$

where w is the lower-right corner element of the $\mathscr{H}^{-1}$ matrix.

In the navigation theory developed here, positional fixes occur at rare intervals. Therefore it seems reasonable, for the purpose of statistical analyses, to postulate the components of the vector $\boldsymbol{\alpha}$ to be a set of statistically independent random variables with zero mean and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_m$. We further assume that between the times T_{n-1} and T_n of two consecutive fixes the clock is drifting at a constant rate, where the rate is random and statistically independent of any previous drift. In other words, if σ_t is the standard deviation in clock drift rate, then the mean squared value of δT_n is postulated to be

$$\overline{(\delta T_n)^2} = \sigma_t^2(T_n - T_{n-1})^2 + \overline{\tau_{n-1}^2} \quad (2.22)$$

We further assume δT to be statistically independent of $\boldsymbol{\alpha}$.

For statistical analyses of navigational errors considered in detail in Section 4, we require the correlation matrix $\mathscr{E}$ and the correlation vector $\boldsymbol{\tau}$ defined by

$$\mathscr{E} = \overline{\boldsymbol{\epsilon}\boldsymbol{\epsilon}'} \quad \boldsymbol{\tau} = \overline{\tau\boldsymbol{\epsilon}} \quad (2.23)$$

Therefore, from equations (2.20) and (2.21) we have

$$\mathscr{E} = \mathscr{Y}\mathscr{S}^{-2}\mathscr{Y}' + \mathscr{X}\mathbf{w}\mathbf{w}'\mathscr{X}'\overline{(\delta T)^2}\sigma_t^4 \quad (2.24)$$

$$\boldsymbol{\tau} = \mathscr{Y}\mathscr{S}^{-2}\mathbf{f} + w\mathscr{X}\mathbf{w}\overline{(\delta T)^2}/\sigma_t^4 \quad (2.25)$$

Finally, the mean-squared error in time estimation is found from

$$\overline{\tau^2} = \mathbf{f}'\mathscr{S}^{-2}\mathbf{f} + w^2\overline{(\delta T)^2}/\sigma_t^4 \quad (2.26)$$

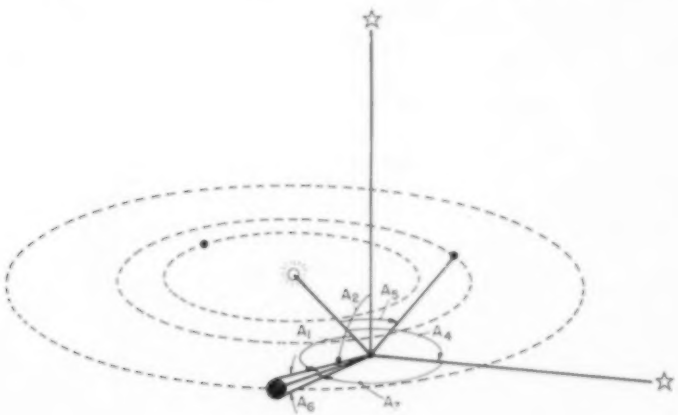


Fig. 3. Measurement of celestial angles.

In our particular studies of the navigation problem, the specific pattern of sightings was arrived at in the following way. Let P_1 and P_2 denote the visible planets which are closest and second closest to the spaceship, respectively. The property of being "visible" in this context means that the line of sight to the planet must lie no closer than some initial angle to the Sun line. The critical value chosen is 15° for the most

part; however, when the proximity of a planet is such that an appreciable disc is displayed, the value is dropped to 3° if need be.

Alpha Centauri, Sirius and Arcturus are among the brightest stars and form very roughly an orthogonal triad. Furthermore, they are so oriented that all three will be visible at any point on the spaceship trajectory according to the 15° criterion.

Now at each fix the measured angles, as illustrated in Fig. 3, are the following: (1) from the Sun to P_1; (2) from Alpha Centauri to P_1; (3) from that one of Sirius or Arcturus to P_1 such that the plane of measurement is the most nearly orthogonal to the plane of the angle measured in (2); (4) from the Sun to the same star selected in (3); (5) from the Sun to P_2, provided that more than one planet is "visible"; and (6) the angular diameter of P_1, provided that it exceeds 1 mrad. It is seen that a minimum of four angles is measured provided at least one planet is visible.

3. COMPUTATION OF VELOCITY CORRECTIONS

Throughout this study we have assumed that our space vehicle does not deviate substantially from a selected reference trajectory. Under midcourse conditions deviations as large as 1 per cent of an astronomical unit (i.e. about 930,000 miles) are about the maximum which could be expected; generally, they would be much smaller. When the spaceship is navigating close to a planet, it is necessary to keep deviations from course to within a per cent or so of the distance to the planet in order to avoid the use of unduly large velocity corrections. The results of the calculations described in Section 5 demonstrate the possibility of achieving this goal. Therefore, it is assumed that perturbation techniques may be used to study these deviations.

Specifically, let $\mathbf{r}_s(t)$ and $\mathbf{v}_s(t)$ denote the position and velocity vectors of the spaceship in an inertial coordinate system with origin at the Sun, and let $\mathbf{g}(\mathbf{r}_s, t)$ denote the gravitational acceleration at position $\mathbf{r}_s$ and time t. Then

$$d\mathbf{r}_s/dt = \mathbf{v}_s \qquad d\mathbf{v}_s/dt = \mathbf{g}(\mathbf{r}_s, t) \tag{3.1}$$

are the basic equations of motion of the spaceship except for those brief periods during which propulsion is applied.

Let the vectors $\mathbf{r}_0(t)$ and $\mathbf{v}_0(t)$ represent the position and velocity at time t associated with the prescribed reference trajectory, and define

$$\delta\mathbf{r}(t) = \mathbf{r}_s(t) - \mathbf{r}_0(t) \quad \delta\mathbf{v}(t) = \mathbf{v}_s(t) - \mathbf{v}_0(t) \tag{3.2}$$

Then, the deviations $\delta\mathbf{r}$ and $\delta\mathbf{v}$ may be approximately related by means of the linearized differential equations

$$d(\delta\mathbf{r})/dt = \delta\mathbf{v} \quad d(\delta\mathbf{v})/dt = \mathscr{R}_0(\mathbf{r}_0, t)\delta\mathbf{r} \tag{3.3}$$

where $\mathscr{R}_0(\mathbf{r}_0, t)$ is a matrix whose elements are the partial derivatives of the components of $\mathbf{g}(\mathbf{r}_0, t)$ with respect to the components of $\mathbf{r}_0$.

A particularly useful fundamental set of solutions of equations (3.3) may be developed in the following way. First, define two matrices $\mathscr{R}(t)$ and $\mathscr{V}(t)$ as the solutions of the matrix differential equations

$$d\mathscr{R}/dt = \mathscr{V} \quad d\mathscr{V}/dt = \mathscr{R}_0\mathscr{R} \tag{3.4}$$

which satisfy the initial conditions

$$\mathscr{R}(T_L) = \mathscr{O} \quad \mathscr{V}(T_L) = \mathscr{I} \tag{3.5}$$

at the time of launch T_L. Here $\mathscr{O}$ and $\mathscr{I}$ denote, respectively, the zero and the identity matrix.

The elements of the $\mathscr{R}$ and $\mathscr{V}$ matrices represent deviations in position and velocity from the reference path as a result of certain specific deviations in the launch velocity from its reference value. For example, the first columns of these matrices are the vector deviations at time t due to a unit change in the first component of the velocity at time T_L.

In a similar manner we may define the matrices $\mathscr{R}^*(t)$ and $\mathscr{V}^*(t)$ as solutions of the same set of matrix differential equations (3.4) which satisfy the terminal conditions

$$\mathscr{R}^*(T_A) = \mathscr{O} \quad \mathscr{V}^*(T_A) = \mathscr{I} \tag{3.6}$$

at time T_A for arrival at the destination planet. Now if we write

$$\delta\mathbf{r}(t) = \mathscr{R}(t)\mathbf{c} + \mathscr{R}^*(t)\mathbf{c}^* \tag{3.7}$$

$$\delta\mathbf{v}(t) = \mathscr{V}(t)\mathbf{c} + \mathscr{V}^*(t)\mathbf{c}^* \tag{3.8}$$

where $\mathbf{c}$ and $\mathbf{c}^*$ are arbitrary constant vectors, it

follows that these expressions satisfy the perturbation differential equations (3.3). Furthermore, they contain precisely the required number of unspecified constants to meet any valid set of initial or boundary conditions. Therefore, equations (3.7) and (3.8) represent the most general solution to the perturbation problem.

Turning now to the navigation problem, we note that since aboard the spaceship only measurements of position are practical, the velocity must, therefore, be inferred by some process. One procedure for obtaining an estimate of spaceship velocity will now be described.

Assume that measured positional deviations $\delta\tilde{\mathbf{r}}_{n-1}$ and $\delta\tilde{\mathbf{r}}_n$ from corresponding reference values are available at the times T_{n-1} and T_n of two successive fixes. Equations (3.7) may then be written twice with T_{n-1} and T_n substituted for t. Therefore, using the notation $\mathscr{R}_n \equiv \mathscr{R}(T_n)$, etc., we have

$$\delta\tilde{\mathbf{r}}_{n-1} = \mathscr{R}_{n-1}\mathbf{c} + \mathscr{R}^*_{n-1}\mathbf{c}^* \quad \delta\tilde{\mathbf{r}}_n = \mathscr{R}_n\mathbf{c} + \mathscr{R}^*_n\mathbf{c}^* \tag{3.9}$$

These equations may be solved for the constant vectors $\mathbf{c}$ and $\mathbf{c}^*$ to yield†

$$\mathbf{c} = \mathscr{R}_{n-1}^{-1}(\mathscr{A}_n^{-1} - \mathscr{A}_n^{*-1})^{-1} (\delta\tilde{\mathbf{r}}_n - \mathscr{R}^*_n\mathscr{R}^{*-1}_{n-1}\delta\tilde{\mathbf{r}}_{n-1}) \tag{3.10}$$

where, for convenience, we have defined

$$\mathscr{A}_n = \mathscr{R}_{n-1}\mathscr{R}_n^{-1} \tag{3.11}$$

Then, from equation (3.8), it follows that‡

$$\delta\tilde{\mathbf{v}}_n^- = \mathscr{V}^-_n\mathbf{c} + \mathscr{V}^{-*}_n\mathbf{c}^* \tag{3.12}$$

Substituting for $\mathbf{c}$ and $\mathbf{c}^*$ from equation (3.10) and its complement, we may write equation (3.12) as

$$\delta\tilde{\mathbf{v}}_n^- = (\mathscr{B}_n + \mathscr{B}^*_n)\delta\tilde{\mathbf{r}}_n + (\mathscr{M}_n + \mathscr{M}^*_n)\delta\tilde{\mathbf{r}}_{n-1} \tag{3.13}$$

† The expression for $\mathbf{c}^*$ may be obtained from equation (3.10) by replacing all starred matrices by the corresponding unstarred ones. This complementary property holds for many of the equations which follow. Therefore, to avoid needless repetition, only one expression will be given. The complementary one may then be had according to the rule stated.

‡ The superscripts $-$ and $+$ are used to distinguish the velocity just prior to a correction from the velocity immediately following the correction.

where, again for convenience, we have defined

$$\mathscr{M}_n = \mathscr{C}_n(\mathscr{A}_n - \mathscr{A}^*_n)^{-1}$$
$$\mathscr{C}_n = \mathscr{V}^-_n\mathscr{R}_n^{-1} \quad \mathscr{B}_n = -\mathscr{M}_n\mathscr{A}^*_n \tag{3.14}$$

Equation (3.13) provides a means of estimating the spaceship velocity at time T_n from positional information at times T_n and T_{n-1}. There remains the problem of determining from these data a required corrective velocity increment $\tilde{\boldsymbol{\Delta}}_n$ which, when added to the calculated velocity deviation $\delta\tilde{\mathbf{v}}_n^-$, will cause the spaceship to arrive at the destination planet. For this purpose let us imagine the target as a point $\mathbf{r}_0(T_A)$ fixed in space and time. Then, if the spaceship arrives at the target from its present position, there will be a velocity deviation $\delta\tilde{\mathbf{v}}(T_A)$ upon arrival which is related to $\delta\tilde{\mathbf{r}}_n$ by the equations

$$\delta\tilde{\mathbf{r}}_n = \mathscr{R}^*_n\delta\tilde{\mathbf{v}}(T_A) \tag{3.15}$$

The corresponding velocity deviation at time T_n

$$\mathscr{V}^{-*}_n\delta\tilde{\mathbf{v}}(T_A) = \mathscr{V}^{-*}_n\mathscr{R}^{*-1}_n\delta\tilde{\mathbf{r}}_n = \mathscr{C}^*_n\delta\tilde{\mathbf{r}}_n \tag{3.16}$$

is precisely that which must be established at time T_n. By combining equations (3.16) and (3.13), we obtain for the required velocity correction

$$\tilde{\boldsymbol{\Delta}}_n = \mathscr{C}^*_n\delta\tilde{\mathbf{r}}_n - \delta\tilde{\mathbf{v}}_n^- = \mathscr{H}_n\delta\tilde{\mathbf{r}}_n - \mathscr{P}_n\delta\tilde{\mathbf{r}}_{n-1} \tag{3.17}$$

where the matrices $\mathscr{H}_n$ and $\mathscr{P}_n$ are defined by

$$\mathscr{H}_n = \mathscr{C}^*_n - (\mathscr{B}_n + \mathscr{B}^*_n) \quad \mathscr{P}_n = \mathscr{M}_n + \mathscr{M}^*_n \tag{3.18}$$

Equation (3.17) is the fundamental guidance equation used by the spaceship computer. Since the matrices involved in these computations depend only upon the reference trajectory and the times of the check-points, they may be considered as pre-calculated quantities. Therefore, when the spaceship computer is presented with two consecutive positional determinations, the computation of the required velocity correction consists of nothing more complex than a few simple multiplications and additions.

Before closing this section we will derive a relationship which will show explicitly how the velocity correction at time T_n is related to the initial velocity errors at launch, the errors associated with the positional fix and the errors

in establishing desired velocity corrections at previous check-points. This will then serve as a basis for the statistical analysis described in Section 4.

For this purpose let $\boldsymbol{\Delta}_n$ and $\boldsymbol{\eta}_n$ denote, respectively, the actual velocity applied at time T_n and the error made in the application of the desired correction $\tilde{\boldsymbol{\Delta}}_n$. Thus

$$\tilde{\boldsymbol{\Delta}}_n = \boldsymbol{\Delta}_n + \boldsymbol{\eta}_n \tag{3.19}$$

Similarly, we define $\boldsymbol{\epsilon}_n$ and $\boldsymbol{\delta}_n$, respectively, as the vector difference between the inferred and actual position and velocity deviations at time T_n, i.e.

$$\delta\tilde{\mathbf{r}}_n = \delta\mathbf{r}_n + \boldsymbol{\epsilon}_n \quad \delta\tilde{\mathbf{v}}_n^- = \delta\mathbf{v}_n^- + \boldsymbol{\delta}_n \tag{3.20}$$

Therefore, it follows from equation (3.13) that

$$\boldsymbol{\delta}_n = (\mathscr{B}_n + \mathscr{B}_n^*)\boldsymbol{\epsilon}_n + (\mathscr{M}_n + \mathscr{M}_n^*)\boldsymbol{\epsilon}_{n-1} \tag{3.21}$$

Then, from equations (3.17) and (3.19), we have

$$\boldsymbol{\Delta}_n = \mathscr{C}_n^*\delta\mathbf{r}_n - \delta\mathbf{v}_n^- + \mathscr{H}_n\boldsymbol{\epsilon}_n - \mathscr{P}_n\boldsymbol{\epsilon}_{n-1} - \boldsymbol{\eta}_n \tag{3.22}$$

It now remains to express $\mathscr{C}_n^*\delta\mathbf{r}_n - \delta\mathbf{v}_n^-$ in terms of the error vectors at the present and previous check-points.

To this end we note that at time T_n^- we have

$$\mathscr{C}_n^*\delta\mathbf{r}_n - \delta\mathbf{v}_n^- = (\mathscr{C}_n^*\mathscr{R}_n - \mathscr{V}_n)\mathbf{c} \tag{3.23}$$

which is obtained by premultiplying equation (3.7) by $\mathscr{C}_n^*$ and subtracting equation (3.8) with $t = T_n^-$. The constant vector $\mathbf{c}$ is determined from equations (3.7) and (3.8) with $t = T_{n-1}^+$. We find

$$\mathbf{c} = (\mathscr{V}_{n-1} - \mathscr{C}_{n-1}^*\mathscr{R}_{n-1})^{-1}(\delta\mathbf{v}_{n-1}^+ - \mathscr{C}_{n-1}^*\delta\mathbf{r}_{n-1}) \tag{3.24}$$

so that by defining

$$\mathscr{D}_n = (\mathscr{V}_n - \mathscr{C}_n^*\mathscr{R}_n)(\mathscr{V}_{n-1} - \mathscr{C}_{n-1}^*\mathscr{R}_{n-1})^{-1} \tag{3.25}$$

and noting that

$$\delta\mathbf{v}_{n-1}^+ = \delta\mathbf{v}_{n-1}^- + \boldsymbol{\Delta}_{n-1} \tag{3.26}$$

equation (3.23) may be written as

$$\mathscr{C}_n^*\delta\mathbf{r}_n - \delta\mathbf{v}_n^- = \mathscr{D}_n(\mathscr{C}_{n-1}^*\delta\mathbf{r}_{n-1} - \delta\mathbf{v}_{n-1}^- - \boldsymbol{\Delta}_{n-1}) \tag{3.27}$$

But equation (3.22) obtains at each check-point so, in particular, at time T_{n-1} we have

$$\boldsymbol{\Delta}_{n-1} + \delta\mathbf{v}_{n-1}^- - \mathscr{C}_{n-1}^*\delta\mathbf{r}_{n-1} = \mathscr{H}_{n-1}\boldsymbol{\epsilon}_{n-1} - \mathscr{P}_{n-1}\boldsymbol{\epsilon}_{n-2} - \boldsymbol{\eta}_{n-1} \tag{3.28}$$

Hence, substituting from equations (3.27) and (3.28), equation (3.22) may be put in the following form:

$$\boldsymbol{\Delta}_n = \mathscr{H}_n\boldsymbol{\epsilon}_n - \mathscr{I}_n\boldsymbol{\epsilon}_{n-1} + \mathscr{G}_n\boldsymbol{\epsilon}_{n-2} - \boldsymbol{\eta}_n + \mathscr{D}_n\boldsymbol{\eta}_{n-1} \tag{3.29}$$

where we have defined

$$\mathscr{I}_n = \mathscr{D}_n\mathscr{H}_{n-1} + \mathscr{P}_n \quad \mathscr{G}_n = \mathscr{D}_n\mathscr{P}_{n-1} \tag{3.30}$$

Equation (3.29) expresses explicitly the velocity impulse actually applied at time T_n in terms of present and previous measurement errors as well as errors in controlling the applied velocity. It is interesting to note, in the case of positional measurements, that errors at the past two check-points contribute to the velocity correction.

For purposes of a statistical analysis the quantity which one would prefer to calculate is the root-mean-square (r.m.s.) value of the sum of the magnitudes of the various velocity changes. However, this cannot be computed without specifying the statistical properties of the errors in considerably greater detail; even then it would be a formidable task. The quantity which we can compute with relative ease is the mean-squared value of the velocity change $\overline{\boldsymbol{\Delta}_n^2}$ at each check-point. If we define the correlation matrix $\mathscr{N}_n$ as

$$\mathscr{N}_n = \overline{\boldsymbol{\eta}_n\boldsymbol{\eta}_n'} \tag{3.31}$$

and assume independence of the sighting errors, then we have

$$\overline{\boldsymbol{\Delta}_n^2} = tr(\mathscr{H}_n\mathscr{E}_n\mathscr{H}_n' + \mathscr{I}_n\mathscr{E}_{n-1}\mathscr{I}_n' + \mathscr{G}_n\mathscr{N}_{n-2}\mathscr{G}_n' + \mathscr{N}_n + \mathscr{D}_n\mathscr{N}_{n-1}\mathscr{D}_n') \tag{3.32}$$

from equation (3.29); where *tr* denotes the trace of a matrix.

4. GUIDANCE ERROR ANALYSIS

In order to study the effect on fuel consumption of errors in guidance, we calculate the mean-squared velocity change $\overline{\boldsymbol{\Delta}_n^2}$ applied at each check-point $n = 1, 2, \ldots, N$. This may be

accomplished using equation (3.32) if we adopt the convention that $\epsilon_0 = \epsilon_{-1} = \epsilon_{-2} = 0$ and agree that $\Delta_0 = -\eta_0$ is to stand for the error in launch velocity.

The correlation matrix $\mathscr{C}_n$, appearing in the equation for $\overline{\Delta_n^2}$, is determined from equation (2.24), with equations (2.22) and (2.26) used to calculate $\overline{(\delta T_n)^2}$ as required. The correlation matrix $\mathscr{N}_n$ is calculated by assuming that η_n is statistically independent of the corresponding Δ_n but, nonetheless, such that its r.m.s. value is a predetermined percentage of the r.m.s. Δ_n. In short, we have tried, within the framework of a rather crude assumption, to take account of the fact that large values of Δ_n will probably produce relatively large errors without studying a detailed mechanism. We further assume η_n to be isotropic statistically so that its correlation matrix $\mathscr{N}_n$ is simply proportional to the identity matrix. In the computation of equation (3.32) all terms except for $tr(\mathscr{N}_n)$ are evaluated first; this term is then obtained as a fixed fraction of the sum of the others.

The computation of $\overline{\Delta_n^2}$ is made first, and the value of $tr(\mathscr{N}_0)$ used in this calculation is the mean-squared value of velocity error in leaving Earth, after Earth's gravity is overcome. The quantity to which a direct value is assigned, however, is the corresponding error at the surface of the Earth, i.e. at burn-out of the main propulsion. Therefore, the magnification factor $[1 + (v_e/v_R)^2]$ is applied to the mean-squared value at the Earth's surface to obtain the corresponding value after escape.

The effect of applying velocity corrections at each of the various check-points will be a velocity deviation from the reference value at the destination planet even if the positional deviation at the target has been successfully nulled. While this velocity error is probably not important on return to Earth, the effect at the target planet may not be ignored. Clearly, an error in planetary contact will result in an error in establishing the spaceship in the proper path for the return trip. We may correct this velocity error in the immediate vicinity of the planet or let the positional error which will result be corrected at the first check-point on the return voyage.

For use in the statistical analysis this velocity deviation $\delta\mathbf{v}(T_A)$ upon arrival at the planet may be expressed in terms of the cumulative effect of velocity corrections at the various check-points. More specifically, let us assume that at time T_n the spaceship, heretofore traveling the reference course, is suddenly given a velocity impulse Δ_n. Our task is to extrapolate the effect of this impulse to the time T_A. Then, under the assumed linearity, we may superpose the results.

For this purpose, we write equations (3.7) and (3.8) for $t = T_n$ with $\delta\mathbf{r}(T_n) = 0$ and $\delta\mathbf{v}(T_n) = \Delta_n$. Solving for $\mathbf{c}$ and $\mathbf{c}^*$ we find

$$\mathbf{c} = (\mathscr{Y}_n - \mathscr{C}_n^* \mathscr{R}_n)^{-1} \Delta_n \tag{4.1}$$

Then, if $\delta\mathbf{v}_n(T_A)$ is the velocity deviation at time T_A due to the velocity impulse at time T_n, it follows from equation (3.8) that

$$\delta\mathbf{v}_n(T_A) = \mathscr{Y}(T_A)\mathbf{c} + \mathbf{c}^* \tag{4.2}$$

(Note that $\mathscr{Y}^*(T_A)$ is the identity matrix.) Now assuming N check-points, the total effect, obtained by superposition, is expressible as

$$\delta\mathbf{v}(T_A) = \sum_{n=0}^{N} [\mathscr{Y}(T_A) - \mathscr{R}_n^{*-1} \mathscr{R}_n](\mathscr{Y}_n - \mathscr{C}_n^* \mathscr{R}_n)^{-1} \Delta_n \tag{4.3}$$

where Δ_0 is the initial velocity error at launch. By means of equation (3.29) $\delta\mathbf{v}(T_A)$ may be expressed in terms of the errors ϵ and η and the mean-squared value computed in the same manner as equation (3.32) was obtained.

Turning now to the problem of guidance accuracy, we may determine the positional deviation at the time of arrival by using equations (3.7) and (3.24) in the form

$$\delta\mathbf{r}(T_A) = \mathscr{R}(T_A)(\mathscr{Y}_N - \mathscr{C}_N^* \mathscr{R}_N)^{-1} (\delta\mathbf{v}_N^+ - \mathscr{C}_N^* \delta\mathbf{r}_N) \tag{4.4}$$

(Note that $\mathrm{R}^*(T_A) = \mathscr{O}$.) Then using equations (3.22) and 3.26), we obtain

$$\delta\mathbf{r}(T_A) = \mathscr{R}(T_A)(\mathscr{Y}_N - \mathscr{C}_N^* \mathscr{R}_N)^{-1}(\mathscr{H}_N \epsilon_N - \mathscr{P}_N \epsilon_{N-1} - \eta_N) \tag{4.5}$$

Only the component of $\delta\mathbf{r}(T_A)$ perpendicular to the direction of relative motion between

vehicle and the target planet is of interest in determining the actual miss distance. The other component along the direction of motion is more nearly responsible for a wrong time of arrival at the planet. Denoting the miss distance vector by $\delta\mathbf{r}_a$ and the velocity vectors of the spaceship relative to the planet at time T_A by $\mathbf{v}_R(T_A)$, we have

$$\delta\mathbf{r}_a = [\mathscr{I} - \mathbf{v}_R(T_A)\mathbf{v}_R'(T_A)/v_R(T_A)^2]\delta\mathbf{r}(T_A) \quad (4.6)$$

Computation of the mean-squared value is accomplished as before.

When the component along the relative velocity vector of the vector difference between the spaceship and planet positions is zero, the spaceship will have reached what we term the "target plane". From our previous discussion it is clear that the time of arrival at the target plane does not coincide with the time T_A. However, if we define δT_A^* as the difference between the actual time that the spaceship penetrates the target plane and the time T_A, then δT_A^* may be computed from the requirement that the vector $\delta\mathbf{r}(T_A) + \mathbf{v}_R(T_A)\delta T_A^*$ be normal to the relative velocity vector $\mathbf{v}_R(T_A)$. Thus we have

$$\delta T_A^* = -\mathbf{v}_R'(T_A)\delta\mathbf{r}(T_A)/v_R(T_A)^2 \quad (4.7)$$

Because of errors in the measurement of time, the time increment δT_A^* does not represent the entire error in the estimate of arrival time at the target plane. At the last check-point $t = T_N$, the error in the determination of the correct time is τ_N. If we assume no drift in the clock during the period from the last fix until final contact with the planet, then when the clock indicates the time T_A, the actual time will be $T_A + \tau_N$. At this time the deviation in position will be $\delta\mathbf{r}(T_A) + \mathbf{v}_R(T_A)\tau_N$ so that by defining δT_{EA} as the error in the estimated time of arrival at the target plane, we have

$$\delta T_{EA} = \delta T_A^* - \tau_N \quad (4.8)$$

Again it is a straightforward process to compute $(\delta T_{EA})^2$ using equations (4.7) and (4.5). Note, however, that the correlation vectors $\boldsymbol{\tau}_N$ and $\boldsymbol{\tau}_{N-1}$, defined by equation (2.23), will be involved in the result.

The error δT_{EA} is the amount by which the spaceship computer would fail to predict the time interval from the last check-point to arrival at the planet. It is of interest at Mars or Venus because it represents the error in timing that would occur for the acquisition of scientific data without a final sighting on the planet during contact. Furthermore, the error is of interest upon return to Earth because it represents a source of East–West impact error due to Earth's rotation.

In our treatment of the round-trip problem we have considered the outbound and return portions of the voyage independently with the sole connecting link between the two being the isotropic random initial velocity error for the return trip. This we have assumed as arising from the effect of gravity of the target planet coupled with a position error of the vehicle. This error provides a $\overline{\eta_0^2}$ for the return trip in the same manner as does the launch velocity error on the outbound trip. Further, it is assumed, as in the case of launch from Earth, that the return trip starts with a perfect knowledge of position and time. If we postulate a fix very near the planet, then this assumption is substantially correct. The lack of knowledge of the correct time is, typically, about 4 min leading to a corresponding uncertainty in position due to the planet's velocity.

To calculate $\overline{\eta_0^2}$ for the return trip, we note that the only effect of the planet's gravitational field during contact is a simple rotation of the relative velocity vector $\mathbf{v}_R$ through an angle 2γ where

$$\tan\gamma = \mu_p/r_a v_R^2 \quad (4.9)$$

as shown in Fig. 4. Here μ_p denotes the gravitational constant of the target planet, and r_a is the distance from the center of the planet to the point in the plane normal to the relative velocity vector at which this vector is pointed from a great distance. Now the magnitude of the difference between the outbound and inbound relative velocity vectors is simply $2v_R \sin\gamma$ so that

$$\overline{\eta_0^2} = (v_R \cos\gamma \sin 2\gamma/r_a)^2\overline{(\delta\mathbf{r}_a)^2} \quad (4.10)$$

where $\overline{(\delta\mathbf{r}_a)^2}$ is obtained using equation (4.6).

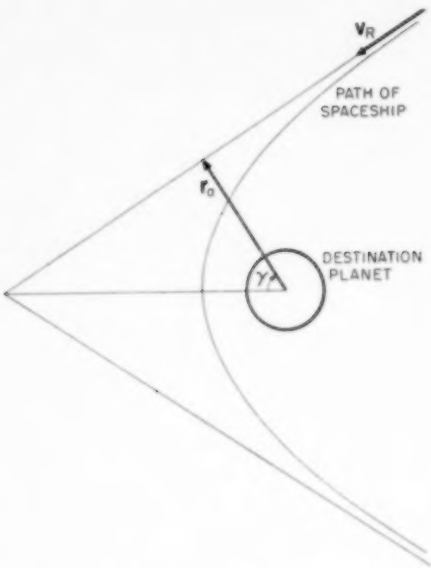


Fig. 4. Motion of spaceship in vicinity of target planet.

Admittedly, this is somewhat of an over-simplification, and more elaborate mechanisms have been devised. However, equation (4.10) was chosen primarily for ease of programming in our computer studies.

5. COMPUTATION RESULTS AND CONCLUSIONS

As noted in the Introduction, a computational procedure for determining three-dimensional round-trip trajectories exists and has been used to make a partial survey of the various possible free-fall routes to the planets Venus and Mars with launch dates in the years 1962–63. Three of these trajectories were selected for use as samples in analysing the navigational scheme described in Sections 2 and 3 and are illustrated in Fig. 5.

The illustrations show the orbits of the spaceship and the planets Venus, Earth and Mars. The paths are shown as solid lines when the orbital plane is above the plane of the ecliptic and broken lines when below. The launch, arrival and return positions are marked with the corresponding dates. The configuration of the spaceship and the planets is shown once during both the outbound and the return portions of the trip. The particular dates chosen coincide with the times for making a celestial fix and an associated velocity correction as determined from the computations to be described later in this section. Of course, more fixes are taken than are actually illustrated in the figures.

For the Venus mission the vehicle velocity relative to the Earth after escape from the Earth's gravitational field is 15,000 ft/sec. After 0·3940 years the spaceship passes 5932 miles from the surface of the planet with a relative velocity of 25,100 ft/sec and returns to Earth 0·8635 years later, entering the atmosphere with a velocity of 50,738 ft/sec. For the Mars trajectory No. 1 the relative velocity at departure is also 15,000 ft/sec, and the vehicle makes one complete orbit of the Sun before contacting Mars 2·4223 years after launch. The ship passes 4693 miles from the Martian surface with a relative velocity of 11,390 ft/sec and returns to Earth 0·8006 years later with a re-entry velocity of 43,189 ft/sec. The departure velocity for the Mars trajectory No. 2 is 12,000 ft/sec, and 1·1970 years is consumed for the outbound trip. After passing 4903 miles from the surface with a relative velocity of 21,657 ft/sec, the vehicle makes one complete orbit of the Sun and returns to Earth 1·9131 years after contact with a re-entry velocity of 39,921 ft/sec.

Before describing our computer studies of the navigation problem, let us briefly summarize the steps employed by the vehicle on the outbound portion of the trip. When the spaceship clock indicates time T_1 (in our numerical studies T_1 is generally 8 hr after launch), a fix is taken and the data is processed by the techniques described in Section 2 to give a calculated value for $\delta\tilde{\mathbf{r}}_1$. Taking $\delta\tilde{\mathbf{r}}_0$ to be zero (or including an allowance for motion of the Earth during a fraction of an hour hold during count-down at launch), equation (3.17) is evaluated for $n = 1$ to give the desired velocity correction vector $\tilde{\mathbf{\Delta}}_1$. The matrices $\mathscr{M}_1$ and $\mathscr{P}_1$ are stored constant data associated with the first check-point. The required correction is then made by the micro-rocket system under control of the computer. The effect of this first correction is to remove the major part of the launch velocity

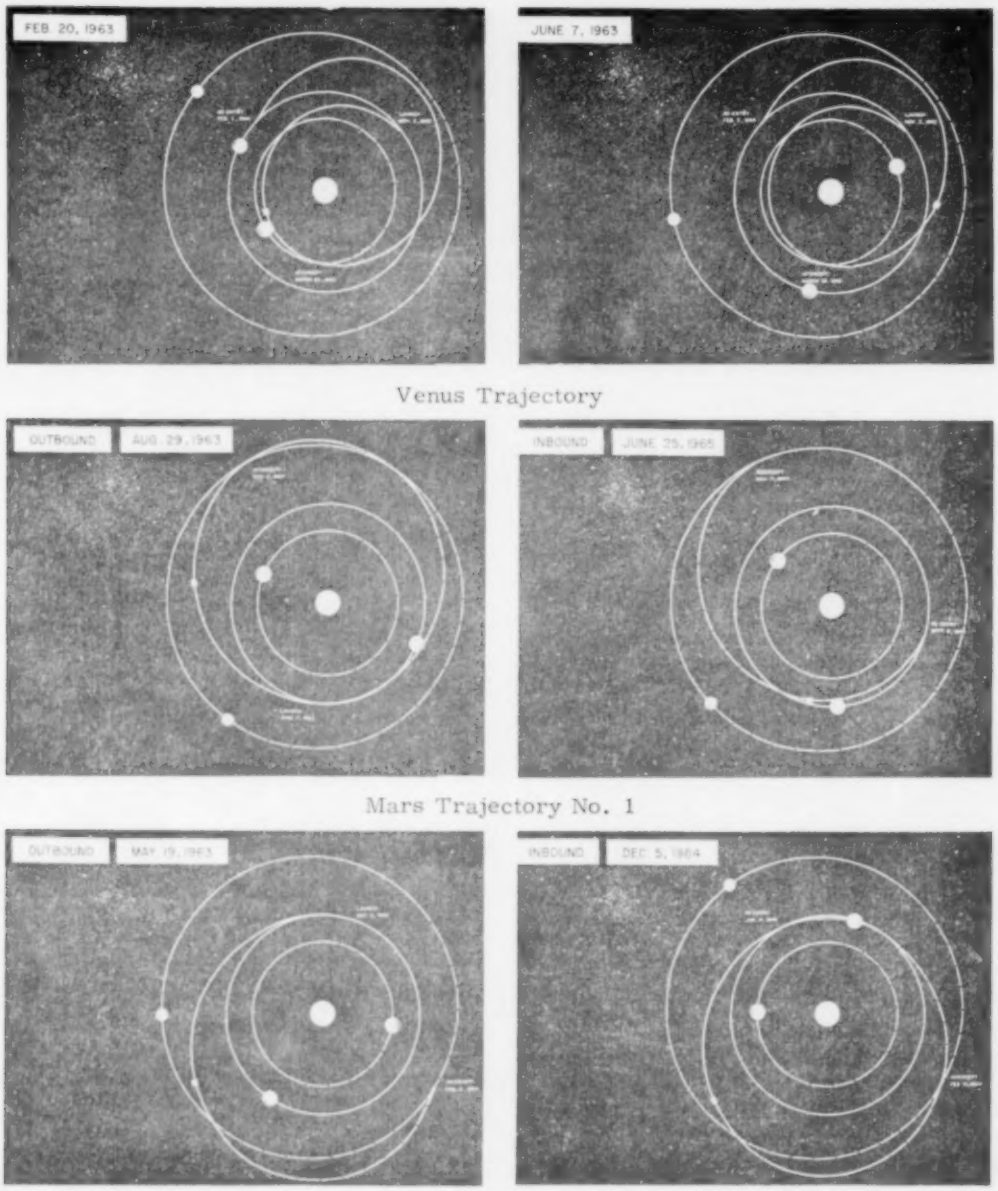


Fig. 5. Sample round-trip trajectories.

error, provided that the fix and resulting velocity correction are made with a fairly good degree of accuracy.

At time T_2 a second fix is taken, leading to a value of $\delta\vec{\mathbf{r}}_2$. Using $\delta\vec{\mathbf{r}}_2$ and the value $\delta\vec{\mathbf{r}}_1$ saved from the first fix, together with stored values for $\mathscr{K}_2$ and $\mathscr{P}_2$, a new velocity correction $\tilde{\boldsymbol{\Delta}}_2$ is computed. The process is repeated for subsequent check-points until the last correction $\tilde{\boldsymbol{\Delta}}_N$ is made. At this time the vehicle is presumably a few hundred thousand miles from the destination planet and established in a satisfactory

passing trajectory. Then at a suitable time following planetary contact, the first of a new series of check-points is initiated for the return trip and the above process repeated until arrival back at Earth.

fix in enhanced. On the other hand, with only a relatively small amount of time left to correct for a positional error, it can happen that an inordinately large velocity impulse will be called for. However, if one does a careful job of

Table 1. One-fix Data

Venus trajectory				Mars trajectory No. 1				Mars trajectory No. 2			
Time (yrs. from depart.)	R.M.S. sight error (miles)	R.M.S. vel. cor. (ft/sec)	R.M.S. miss dist. (miles)	Time (yrs. from depart.)	R.M.S. sight error (miles)	R.M.S. vel. cor. (ft/sec)	R.M.S. miss dist. (miles)	Time (yrs. from depart.)	R.M.S. sight error (miles)	R.M.S. vel. cor. (ft/sec)	R.M.S. miss dist. (miles)
Out-bound				Out-bound				Out-bound			
0·00114	80	11	78100	0·00114	66	10	315810	0·00125	51	7	216649
0·00200	203	17	124858	0·050	6118	23	657472	0·100	4109	10	229397
0·025	7892	63	385808	0·100	4777	17	260057	0·200	7192	72	127212
0·100	1826	7	19387	0·200	4848	9	153416	0·450	10175	5	24678
0·120	1952	20	16108	0·300	10502	64	214724	0·600	9379	6	17088
0·122	1963	43	15793	0·400	5988	4	67901	0·800	10991	8	15469
0·124	1972	685	15464	0·600	9488	9	40102	1·000	8372	5	3778
0·125	1975	79	15299	0·650	11565	33	56263	1·18800	11054	203	74
0·126	1979	44	15134	0·675	13052	398	481285	1·1970†			
0·200	2339	5	7292	0·700	13568	36	49078	Return			
0·300	3763	7	3990	0·875	15086	72	21858	0·00100	199	33	303904
0·365	5818	33	1605	0·900	14685	1276	21283	0·00274	1478	90	823173
0·375	8449	74	180	0·925	13909	65	18741	0·300	7719	9	49609
0·38900	3488	117	46	1·200	10193	5	6100	0·425	9533	106	70175
0·3940†				1·400	8317	5	18631	0·435	8444	308	48283
Return				1·500	13550	56	180764	0·440	8482	680	48411
0·00100	149	25	507577	1·525	14248	1193	3556913	0·445	8533	167	48574
0·00150	333	38	759809	1·550	15092	65	175835	0·450	8594	97	48772
0·025	7131	58	974148	2·050	7827	5	10786	0·500	10393	48	59566
0·050	6440	31	419326	2·175	10412	27	40764	0·520	11004	206	111854
0·100	8467	38	232021	2·200	11628	127	149689	0·530	11349	160	89111
0·200	7301	10	76838	2·225	12867	24	21316	0·540	11721	54	60156
0·300	12458	14	80446	2·41000	6014	82	54	0·800	10924	2	17339
0·400	12668	12	50550	2·4223†				1·000	19440	5	20840
0·500	10944	10	26887	Return				1·300	10233	126	28759
0·600	9363	10	15485	0·00100	55	10	138531	1·400	10101	18	81272
0·700	7651	19	16097	0·200	7829	14	88619	1·500	10670	77	284620
0·750	9625	15	5678	0·400	6671	8	22231	1·800	7990	9	2942
0·800	11003	29	3520	0·600	4067	3	3606	1·90000	4747	54	81
0·85800	7264	221	73	0·79600	2654	96	40	1·90800	1462	46	30
0·8635†				0·8006†				1·9130†			

† Arrival time.

It is fundamental to our theory that navigation in the vicinity of a planet is treated in exactly the same manner as the mid-course problem. Being in proximity to the target has two effects which are somewhat conflicting considering the objectives of the navigation system. On the one hand, when the spaceship is near a celestial object, the accuracy in obtaining a positional

selecting check-points during mid-course flight, it is possible to keep these final velocity corrections to within reasonable size.

For the purpose of selecting suitable check-points, a number of one-fix runs were studied. In this type of calculation it is assumed that the spaceship leaves Earth with no initial velocity error. At the designated time for the single fix

a velocity change is introduced based on the erroneous determination of position. As a result it misses the target planet by an amount whose r.m.s. value is given. Some of the results from these one-fix calculations are summarized in Table 1. Noting the frequent fluctuations in as the square root of the trace of the $\mathcal{E}$ matrix; (3) the r.m.s. velocity correction in feet per second as calculated from equation (3.32); and (4) the r.m.s. miss distance at the target planet in miles as determined using equation (4.6).

The time intervals were chosen to be some-

Table 2. Round Trip Data Venus Trajectory

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec)
0·00114	66	107	0·000			
0·3000*	3762	138	0·024			
0·3890	3432	210	0·022			
0·3940	—	228	—	0·213	165	163
Tot. vel. cor. = 683 ft/sec						
0·00100	149	166	0·000			
0·2000*	7301	68	0·017			
0·7500	9621	120	0·047			
0·8580	6060	489	0·041			
0·8635	—	—	—	0·265	245	
Tot. vel. cor. = 1526 ft/sec.						
0·00114	66	107	0·000			
0·1750	2181	41	0·015			
0·3500	4547	43	0·020			
0·3875	5005	237	0·019			
0·3940	—	221	—	0·349	339	336
Tot. vel. cor. = 650 ft/sec						
0·00100	149	338	0·000			
0·4000	12667	137	0·034			
0·7500	9620	86	0·043			
0·8560	7528	435	0·038			
0·8635	—	—	—	0·339	388	
Tot. vel. cor. = 1645 ft/sec						

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec)
0·00114	66	107	0·000			
0·1750	2181	41	0·015			
0·3600	5019	56	0·021			
0·3875	5011	252	0·019			
0·3940	—	232	—	0·368	386	383
Tot. vel. cor. = 688 ft/sec						
0·00100	149	384	0·000			
0·6000	9362	306	0·053			
0·7750	9329	49	0·049			
0·8560	7542	455	0·042			
0·8635	—	—	—	0·346	420	
Tot. Vel. cor. = 1882 ft/sec						
0·00114	66	107	0·000			
0·3000*	3762	138	0·024			
0·3850	6918	178	0·022			
0·3940	—	203	—	0·463	492	487
Tot. vel. cor. = 627 ft/sec						
0·00100	149	489	0·000			
0·3000	12457	98	0·026			
0·7500	9620	127	0·044			
0·8560	7530	413	0·039			
0·8635	—	—	—	0·339	383	
Tot. vel. cor. = 1753 ft/sec						

the data, one is easily convinced that the problem of selecting optimum times of fixes is by no means one possessing a simple and obvious solution.

The quantities tabulated are: (1) the time of the fix in years of travel from the departure planet; (2) the r.m.s. sighting error in miles computed

what irregular in an effort to give better definition to regions in which rapid changes are occurring. One notes that sudden sharp peaks in velocity correction occur from time to time, unaccompanied by any marked change in sighting errors, e.g. the points $t = 0{\cdot}440$ and $0{\cdot}520$ for Mars trajectory No. 2 return. The phenomenon is

explained by the fact that these times correspond to points on the spaceship trajectory which are an integral multiple of 180° from the launch or destination position. The explanation lies in the fact that at an arbitrary point S a velocity change in the direction normal to the plane of the orbit has the effect of rotating the orbital plane about the line from the Sun to S, but leaves positions unchanged at points 180°, 360°, etc., away from S. Consider, for example, the point $t = 0{\cdot}520$; the spaceship has travelled 178° since departing from Mars. Through observational error, a position displacement is computed which has a component of several thousand miles in the direction normal to the plane of motion. A displacement of this magnitude in any other direction is perfectly reasonable; in this particular direction, however, it can be produced only by a large velocity error in the same direction upon leaving Mars. As a result, a large velocity change is called for to compensate for this nonexistent error.

Table 3. Round Trip Data Mars Trajectory No. 1

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec)
0·00114	66	107	0·000			
0·6000	9487	28	0·052			
1·2000*	10168	12	0·054			
2·0500	7250	23	0·080			
2·4100	5992	224	0·065			
2·4223	—	222	—	0·797	203	126
Tot. vel. cor. = 615 ft/sec						
0·00100	55	127	0·000			
0·6000*	4065	130	0·049			
0·7960	2590	191	0·041			
0·8006	—	—	—	0·164	80	
Tot. vel. cor. = 1063 ft/sec						

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec)
0·00114	66	107	0·000			
1·2000*	10193	25	0·063			
2·0500	7294	9	0·084			
2·4140	3483	312	0·067			
2·4223	—	309	—	0·454	209	130
Tot. Vel. cor. = 762 ft/sec						
0·45000	4545	497	0·039			
0·7600	6191	83	0·041			
0·7920	7446	248	0·036			
0·8006	—	—	—	0·621	379	
Tot. vel. cor. = 1589 ft/sec						

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec)
0·00114	66	107	0·000			
1·2000*	10193	25	0·063			
1·9500	7748	9	0·091			
2·4100	5996	325	0·071			
2·4223	—	320	—	0·788	336	209
Tot. vel. cor. = 787 ft/sec						
0·00100	55	209	0·000			
0·4500	4543	59	0·039			
0·7400	5916	36	0·041			
0·7920	7446	230	0·036			
0·8006	—	—	—	0·569	310	
Tot. vel. cor. = 1321 ft/sec						

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec)
0·00114	66	107	0·000			
0·4000	5988	19	0·033			
1·2000*	10171	8	0·056			
2·0500	7256	14	0 081			
2·4100	5993	224	0·066			
2·4223	—	222	—	0·797	203	126
Tot. vel. cor. = 593 ft/sec						
0·25000	8249	770	0·022			
0·6000*	3711	76	0·036			
0·7600	6101	49	0·035			
0·7920	7338	261	0·032			
0·8006	—	—	—	0·614	377	
Tot. vel. cor. = 1749 ft/sec						

In Tables 2, 3 and 4 are summarized the results of several multifix runs for the three trajectories studied. The postulated errors are: (1) an r.m.s. initial velocity error of 40 ft/sec at sea level,

produced by a launch guidance error of 0·1 per cent; (2) an r.m.s. error in applying any desired velocity change of 1 per cent; (3) an r.m.s. clock drift rate of one part in 100,000; (4) an r.m.s. error in angle measurements of 0·05 mrad or 10·3″ of arc. The quantities tabulated are: (1) the time of the fix in years from departure; (2) the r.m.s. sighting error in miles; (3) the r.m.s. velocity correction in ft/sec; (4) the r.m.s. time error in hours as computed from equation (2.26); (5) the r.m.s. error in the estimated time of arrival (ETA) in hours as obtained using equation (4.8); (6) the r.m.s. miss distance at the target planet in miles; and (7) the r.m.s. initial velocity error for the return trip as obtained using equation (4.10). An asterisk, after the time fix in these tables, marks those times for which the configurations are illustrated in Fig. 5.

The runs in these various tables are not optimized, but do indicate the ranges of values which one may reasonably expect to attain. In addition, a number of other runs were made (not included in the tables) in which larger errors were assumed. The following facts are apparent directly from the data or by a reasonable extrapolation:

Table 4. Round Trip Data Mars Trajectory No. 2

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec)
0·00114	43	130	0·000			
0·4500*	10174	29	0·039			
1·0000	8211	24	0·053			
1·1880	10422	319	0·051			
1·1970	—	309	—	0·778	279	127
Tot. vel. cor. 811 = ft/sec						
0·8000*	10924	113	0·067			
1·8000	7518	50	0·076			
1·9000	4744	132	0·040			
1·9080	1420	301	0·031			
1·9131	—	—	—	0·212	90	
Tot. vel. cor. — 1406 ft/sec						
0·00114	43	130	0·000			
1·0000	8372	169	0·065			
1·1880	10597	251	0·061			
1·1970	—	280	—	0·795	283	129
Tot. vel. cor. = 831 ft/sec						
0·00050	50	130	0·000			
0·8000*	10924	13	0·067			
1·8000	7518	40	0·076			
1·9000	4744	132	0·040			
1·9090	965	338	0·031			
1·9131	—	—	—	0·123	90	
Tot. vel. cor. = 1484 ft/sec						

Time of fix (yrs.)	Sight error (miles)	Vel. cor. (ft/sec)	Time error (hr)	ETA error (hr)	Miss dist. (miles)	Vel. err. (ft/sec
0·00114	43	130	0·000			
0·5000	9984	33	0·044			
1·0000	8209	21	0·052			
1·1900	7820	379	0·050			
1·1970	—	367	—	0·572	193	88
Tot. vel. cor. = 930 ft/sec						
0·00050	50	89	0·000			
0·8000*	10924	13	0·067			
1·8000	7518	40	0·076			
1·9000	4744	132	0·040			
1·9090	965	338	0·031			
1·9131	—	—	—	0·123	90	
Tot. vel. cor. = 1542 ft/sec						
0·00114	43	130	0·000			
1·0250	6874	194	0·066			
1·1880	10591	242	0·060			
1·1970	—	285	—	0·800	290	132
Tot. vel. cor. = 851 ft/sec						
1·00000	19440	162	0·087			
1·8000	7538	100	0·076			
1·9000	4744	574	0·041			
1·9090	965	338	0·031			
1·9131	—	—	—	0·125	76	
Tot. vel. cor. = 2025 ft/sec						

(1) Navigation with a position error not in excess of 300 miles r.m.s. at the destination planet or 250 miles r.m.s. upon return to Earth is feasible, with the standard assumed errors, and with the expenditure of no more than 1500 ft/sec r.m.s. velocity change.

(2) Errors in launch velocity and in measurement and control of velocity changes can be substantially degraded (by a factor of 2·5) at the expense of about 30 per cent more fuel.

(3) A high degree of clock accuracy appears to be fairly unimportant. A degradation to 2·5 parts in 100,000 is compensated by less than 4 per cent increase in fuel consumption. At the end of the journey the spaceship computer appears to be telling time mainly by the motion of the planets.

(4) Stated in another way, guidance depends primarily upon the accuracy of optics, for the range of parameter values considered here. Therefore, the way to conserve fuel is by improving the optical accuracy.

REFERENCES

1. J. H. LANING, Jr., E. J. FREY and M. B. TRAGESER, *Vistas in Astronautics*, p. 63, Vol. 2, Pergamon Press, New York (1959).
2. R. H. BATTIN, *J. Aero Space Sci.* (Sept. 1959).

Note added in press: Subsequent to the writing of this paper a numerical analysis has shown that the miss distances quoted in Tables 2, 3, and 4 are in error. A more careful calculation showed that the tabulated numbers are in most cases two or three times larger than the correct values.

A HIGH PRECISION STELLAR NAVIGATOR FOR INTERPLANETARY GUIDANCE

CHARLES D. BOCK
Arma Division, American Bosch Arma Corporation

Abstract—This stellar navigator uses the star studded celestial sphere as its calibrated angular scale. Since the star positions are known to 0·01″ of arc, this new class of instruments promises a substantial improvement in the precision of a fix over conventional star trackers.

The proposed instrument centers its axis on a suitable planet. It then measures the angular distance to a star nearby on the celestial sphere, establishing a line of position. A second planet, well off this line, gives a second line of position. The intersection of these lines is the fix in space.

The need for high precision navigation data on an interplanetary flight has been pointed out by Stearns and Frye.[1] In this study they estimate that data taken with a sextant of 0·5′ accuracy will require fuel expenditure of 10 per cent of the gross weight of the interplanetary vehicle. They calculate that a 1° course change takes about 20 per cent of the vehicle weight, since the velocity includes the Earth's orbital velocity. With presently available fuels much will be gained by more precision.

Even the problem of getting off to a good start is somewhat complicated by the loss of escape velocity. The errors in residual velocity are magnified by the ratio of $V_{\text{total}}/V_{\text{residual}}$. For a Hohman ellipse transfer to Venus or Mars, this factor is about 10×. A table of miss distances corresponding to 1 ft/sec of velocity error is given in the space handbook prepared by the Rand Corporation[2] from data reported by Ehricke.[3]

Table 1.

Destination	Nautical miles
Moon	20–100
Venus	25,000
Mars	20,000
Jupiter	65,000

This error can be relieved by remaining in a satellite orbit long enough to establish a new set of initial conditions for an inertial system, or, if one can afford the necessary network of ground stations, a radio guidance system capable of measuring near the limit of the Earth's gravity field will give improved initial velocity control. Either of these methods is costly in the ground installations.

In any case, it is likely that a mid-course navigation system would be desirable. This system should be as simple as possible and offer the highest precision. It appears that an important development in the history of astronomy may be the key to such a system. It will enable precise establishment of lines in space to suitable planets.

From the time of acceptance of the Copernican theory, many astronomers attempted to measure the distance to the nearer stars using the diameter of the Earth's orbit as a base line. Although the attempts led to several interesting discoveries, no one, not even Sir John Herschel, was successful in measuring this fractional second of angle. In the year 1838, three astronomers used the relative shift of the stars with respect to the positions of distant stars nearby on the celestial sphere. Bessel[4] found the annual parallax of 61 Cygni to be 0·31″. Henderson measured α Centauri and Struve α Lyrae in the same year. They used the heliometer, a sort of split telescope designed by Dolland[5] in 1754, in which two semicircular halves of the objective could be moved relative to each other with a micrometer screw until the images of two stars are made to coincide.

At present, use is made of a photographic technique. By this method, under observatory conditions, a probable error of 0·01″ is obtainable and some of this error arises from the need to separate out the proper motions from the annual parallax motion. Most of the earlier work of the 19th century was carried out with 4 in.–6 in. objectives, which suffice to give better than 0·1″. In modern work, the star image is rarely less than 2″ diameter, but still yields position accuracies of 0·01″ because of the symmetric nature of the aberrations.

Only one mention of this technique has been seen in the literature on interplanetary guidance in a paper by Porter[6] of the British Interplanetary Society. It was thought worthwhile to look at the design problems of instrumentation using this principle.

It is difficult to design an astro-tracker to measure large angles to an accuracy of 1″. The reference circles must have errors the order of a few parts in a million. Either the stable reference must be held to fractional seconds or the telescopes arranged to measure several objects at once. Stuhlinger[7] suggests a setup with ten telescopes to cover five stars, the sun and four planets. Present star trackers have errors measured in minutes rather than seconds. While these problems may not be insoluble, the more conventional star tracker systems do not promise the desired accuracy in a simple inexpensive instrument.

By contrast, use of the already calibrated celestial sphere offers a natural solution to the measurement of large angles in space. The stars may be considered as index markers in the sky, whose position is known to about 0·01″. In the simplest case, occultation of the star by the planet in question would serve to establish the direction of the line from observer to planet. Actual occultation is too rare for navigation use, but a near star can serve as well if we can measure the angular deviation from occultation. The large number of stars assures that there is always one of the desired nearness, although it may not be very bright.

If we assume a reference star half a degree away, 1″ is only 1 part in 2000, for which conventional gearing, or micrometers, are easily made. Let us look at some of the problems in the design of a differential sextant suitable for this task.

If we have a manned vehicle with an officer-navigator, our problem is reduced to one of taking a photograph and supplying the grade of stabilization needed to keep the star images adequately sharp for measurement on the comparator microscope. Added problems are presented if we make this navigation instrument automatic.

The question of what magnitudes of star we must be able to use determines the needed sensitivity. A look at the space distribution of stars of various magnitudes shows that to keep above a probable density of one star per square degree we must be able to use stars as weak as 10th magnitude. The stars are thinnest near the poles of our galaxy, with a distribution as shown in Table 2. Since the galactic pole is about 30° from the ecliptic plane, we must operate near minimum star densities (see Fig. 1).

Table 2. Star Densities

Up to magnitude	Star densities			
4	0·016	0·007	0·005	0·005
6	0·129	0·062	0·042	0·037
8	1·0	0·48	0·32	0·275
10	7·8	3·5	2·2	1·8
	Galactic latitude			
	0°	30°	60°	90°

The light needed for tracking an 8th magnitude star can be collected by a 4 in. telescope provided a good photo-multiplier tube is used. Such a star gives 10^{-14} W/in² at the Earth. An effective aperture of 1 in² would just meet the 1 c/s threshold[8] of the 1*P*21 photo-multiplier. Hence, reasonably good tracking should be possible with an 8-in. aperture at 10th magnitude. By using extra aperture and bandwidth, the need for special cooling of the sensor can be eliminated.

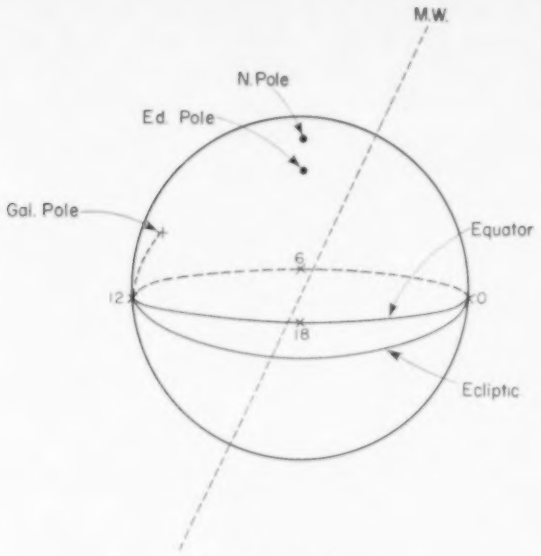


Fig. 1. Pole of galaxy.

The telescope design seems to present no unconventional problems. To keep it compact, it can be made a reflector type, about F 1·2. In the Cassegrain form the effective focal length would be about 30 in. with an over-all length of 10 in. (Fig. 2). In Newtonian form, the assembly could be made shorter and more nearly spherical but may suffer some from asymmetry. A Schmidt corrector plate will be necessary in either form to keep good resolution off the axis, since aberrations are excessive even at $\frac{1}{2}^\circ$. The design should allow a useful field of 5° diameter which is large enough that there will almost always be the needed two stars in the field. The sensors must be mounted to follow the curvature of the surface of focus.

It does not seem that an instrument of the heliometer type would be good here since difficulties occur in superposing stars that differ in brightness. Our problem is to measure the angular position of the planets, giving a serious contrast in illumination.

The temperature compensation of magnification and focus must be given special attention. If good results at 2° are desired, scaling provision will be needed in addition to compensation. The stabilization requirements are severe. Movement around the mean pointing direction should be kept to a few seconds, if the mean value is to be correct to a fraction of a second. This could be managed by making the inertial platform

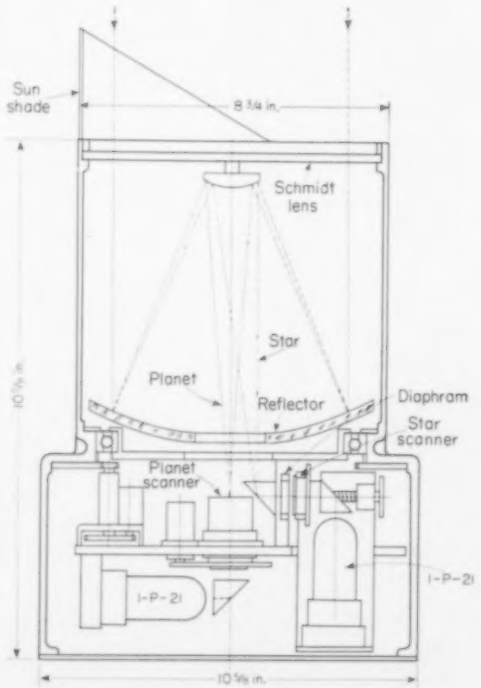


Fig. 2.

integral with the telescope. But a more flexible arrangement is proposed here. A dual gimbal system will be used, with inertial stabilization data holding the outer gimbal on the expected position of the planet and an inner gimbal set to servo from this steady reference by means of

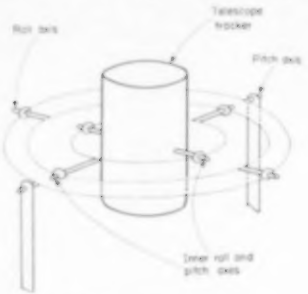


Fig. 3. Telescope mount.

the optical centering signal from the planet sensor. Hence, no high speed requirement is necessary from this optical servo, and it can be made slow to better average out background photocell noise (Fig. 3).

The arrangement of the sensors is shown in Fig. 2. The planet sensor is fixed at the axis of the telescope. It consists of a rotating scanner plate with a generator attached to give reference voltage for phase discrimination, an iris diaphragm to allow acquisition, then masking of the planet from stray stars, and a photomultiplier tube to detect the intensity oscillation due to the scanner-off center planet combination. The crescent shape of the planet image requires a correction since the system will null on the center of brightness.

The star seeker unit is made to allow closest approach to the planet. Hence, the prism and right-angle bends since the scanner needs space for its rotating gear and bearings. It also has an iris diaphragm. Either one or two star seeker units can be used; with two, better scaling data can be obtained with simultaneous use of these. The program can be set up to avoid interference between the two star seekers.

The geometry of the problem is very simple for this system of navigation. The calibrated celestial sphere allows projection of the reference planet's position, and so defines the line in space on which the vehicle lies (Fig. 4). A second planet is used to establish a second line of position. If there were no error in these lines, they would intersect at the position of the vehicle. Hence, the failure to intersect yields a measure of the consistency of one component of the line of position and may be useful as a test of the reliability of the fix. Redundant information is frequently useful in this manner.

The system is shown in Fig. 5. The telescope is pointed at the planet whose position is to be determined by angular orders generated by the programmer. These orders are made up in the computer of the sum of the expected position of the planet less the heading of the missile as

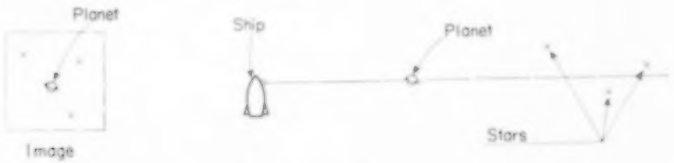


Fig. 4. Line of position.

furnished by the inertial platform. They need only be accurate enough to get light from the planet into the wide open planet scanner. Then the output of the scanner centers the telescope axis on the planet image by means of the inner gimbal servos and holds it there as a fine adjustment to the programmer position. The scanner diaphragm then contracts to its small setting to cut out stray biasing light. This symmetrical masking of the field by the diaphragm also avoids bias, which would arise

from the star tracker on close approach. Any asymmetrical obscuring of the scanner aperture will unbalance the stray light and so shift the centering point.

The star scanner unit is set to acquire the selected star by the programmer which furnishes precalculated values of the expected position in polar coordinates centered on the planet. Distinctive bright stars must be chosen, and a possible search program instituted if the expected light intensity is not observed. Then throughout the voyage, as seen from expected positions on the nominal orbit. This information will be stored in the programmer. At the proper times, the expected positions of the chosen reference planet will be given to the computer, which will add them to the reference position of the vehicle and the inertial reference direction angles to give pointing orders to the telescope. The expected positions of the nearby stars will also be transmitted to the star sensors. The method of acquiring the planet in the telescope

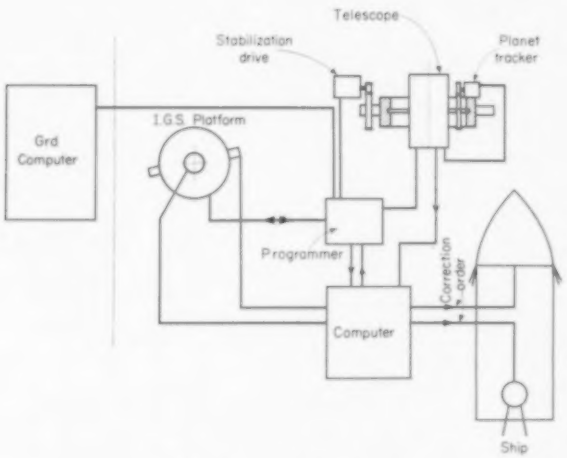


Fig. 5. System functional.

angular deviation of the star from the planet center is read out on the micrometer radius control and circular rack, and this information fed to the computer, which can generate the orders required to correct the missile velocity. The inertial guidance system is useful in precision control of these velocity increments, since it measures the impulse without delays which are inherent in the use of position information, and better accuracy will result from closed loop control of the impulse.

To make full use of differential quantities requires precomputation on the ground. The voyage is planned as it must be to catch the fast moving destination. We will call this the nominal orbit. The apparent position of the planets to be used will be calculated for intervals will depend on design considerations of the whole missile. However, it will not be impractical to rotate the missile to avoid obscuring the planet desired, so easing the mounting problem of the telescope. Note also that no high precision of angular relation between the telescope mounting and the inertial reference is required. We use this relation only for acquisition and can cover a fair fraction of a degree. The accuracy requirement here will be set by the need to correct for gyro drift in the inertial reference.

This is best done on a bright star, and provision should be made for such periodic resetting in the programmer. In the gyro resetting operation, the position of the selected bright star will be stored in the programmer. At intervals based

on the expected drift rate of the platform, the telescope will be directed to this star. Then the telescope will be aligned as in the case of the planet to this star. Two components of the platform error can then be read on the inner gimbal servos and this quantity precessed out of the platform. A second star about 90° from the first can furnish the correction for the third axis of the platform in much the same way, with provision for orthogonality in the precession orders. The computer will generate these orthogonal orders from the observed errors and a repeat of the process will improve the alignment.

After acquisition, the telescope will have its outer gimbals held by the computed position of the planet against motion of the ship, but will be held precisely centered on the planet by the central optical sensor, which will furnish orders to the inner gimbals. This arrangement will relieve the optical link from any high speed response needs, so its performance can be used toward improved precision of pointing.

The measured errors in vehicle position from the nominal position will be fed to the vehicle computer. After a suitable number of these position fixes has been put in, the computer will calculate a correcting increment of velocity. The inertial guidance platform will then monitor the rotation of the ship to the proper direction and meter out the required velocity change.

It should be noted that it would be wasteful of fuel to return to the nominal orbit. Rather, modify the present orbit of the ship to reach the desired objective, since there is an infinity of orbits which will reach the desired objective, and we must keep energy loss to a minimum.

If we use the destination planet as one reference object, we will eliminate errors in its known position, but problems may arise on closer approach when it is too big to fit in the scanning raster.

Let us examine some of the difficulties we must overcome if we are to operate this system to obtain fractional second accuracy. They may be listed as follows, much as Larmore[9] has done:

(1) Planets will appear crescent shaped. Non-predictable variations in albedo and Moon illumination of the dark side will limit the accuracy of correction possible.

(2) Sunlight must be kept off the lenses as the bias would depend on the star magnitude.

(3) Star proper motion must be corrected, along with velocity aberrations and planet position delay.

(4) Double stars may give an average position.

(5) Weaker stars in the field may give a bias, but this can be predetermined if they cannot be masked out.

(6) Relative star magnitudes and background light will differ from that seen on Earth. Use of visual light only will minimize the differences due to atmospheric effects calibration on the Earth.

(7) Long averages of the star position data, perhaps several minutes, will be needed to suppress the sensor noise.

(8) Full symmetry of the optical system must be preserved; e.g. masking part of objective will offset the image too much.

It should be noted that this scheme is of very limited application to earthbound navigation. The problem there is to find the space direction of the gravity vector. If one used the Moon in place of a planet as position reference, the precision needed would increase over the gravity vector case by the ratio of the Moon's distance to the Earth's radius, a 60:1 ratio. Only in case of extreme acceleration conditions could anything be gained, since the angular size of the Moon makes it difficult to center on the Moon to the needed accuracy.

The principles outlined in this paper will permit the design of an automatic differential sextant of accuracy better than 1″ of arc with promise of future development if desired, possibly as fine as 0·1″.

The use of this telescope unit in a navigation system as described would give midcourse

positions in the range of 20–200 miles, very adequate for economical guidance to any of the nearby planets.

REFERENCES

1. —, STEARNS and —, FRYE, *Space Aeronautics*, p. 48 (Mar. 1959).
2. Rand Corporation, *Space Handbook*, U.S. Government Printing Office, Washington, D.C. (1959).
3. K. A. EHRICKE, *Instrumented Comets: Astronautics of Solar and Planetary Probes;* Convair Re/24, AZP-019 (July 1957).
4. F. W. BESSEL, *Astr. Nachr.* 1846, 365.
5. Heliometer, *Encyclopedia Britannica*, **11**, 398 (1957).
6. J. G. PORTER, *J. Brit. Interplanet. Soc.* **13**, 68, (1954).
7. E. STUHLINGER, *Jet Propulsion* **27**, 410 (1957).
8. *American Institute of Physics Handbook*, Vol. 6, p. 114, McGraw-Hill, New York (1958).
9. L. LARMORE, *Aero Space Engineering*, **18**, 37 (Jan. 1959).

A GUIDANCE TECHNIQUE FOR INTERPLANETARY AND LUNAR VEHICLES

WILLIAM SCHROEDER and CLARENCE W. PITTMAN
Space Technology Laboratories, Inc.

Abstract—The problem of guiding an extraterrestrial vehicle during its launch phase by means of a closed loop control system, using a large digital computer in real time, is discussed. Equations are derived from fundamental physics, from which the computer can determine the desired burnout conditions during flight. Steering equations are then derived, by means of which the vehicle will be guided to these burnout conditions in an efficient manner. Compared to the technique of guiding the vehicle to a trajectory determined before the flight, this method has the advantages that it is simple and flexible, and it can accommodate a wider range of vehicle performance perturbations.

The accuracy of the guidance technique is demonstrated by results from simulated flights to Venus.

1. INTRODUCTION

This paper is concerned with the guidance of an extraterrestrial vehicle during its initial powered flight phase. The guidance is accomplished by means of a closed loop control system shown qualitatively in Fig. 1. The objective of the guidance is to direct and terminate the thrust of the engine so that the payload will be injected into a free flight orbit which will take it into the vicinity of some extraterrestrial body.

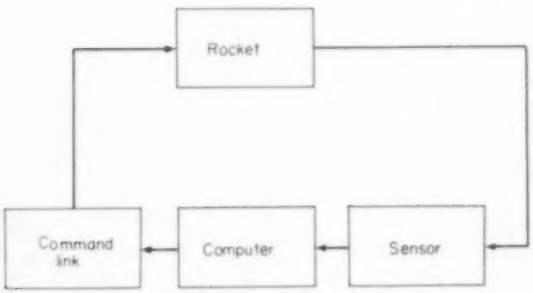


Fig. 1.

The free flight orbit is determined by the position and velocity of the vehicle at thrust termination and the time at which this occurs. One could, in principle, guide the vehicle so that it would arrive at a preselected position and velocity at a preselected time. This approach is not feasible because:

(1) The time of launch would have to be controlled very accurately.

(2) The magnitude of missile thrust would have to be adjustable.

(3) It would be wasteful of fuel to guide to preselected end (cutoff) conditions, without regard to the vehicle's present conditions.

A more desirable approach is to select in the guidance computer a set of end conditions which can be reached efficiently from the missile's present position, velocity and time. These end conditions must, of course, satisfy the guidance objective.

In general, the end conditions (position, velocity and time) must all be specified. But at any position and time there is at least one velocity which satisfies the objective so that it is sufficient to control only velocity. Accepting the end position and time, computing the corresponding required velocity and then steering the vehicle to attain the required velocity is economical of fuel and is readily adaptable to closed loop control.

The exact determination of the required velocity is too complex to be performed in presently available computers in real time. It requires the solution of an n-body problem, for which no explicit solution exists. An approximate solution can be found in either of two ways:

(1) Expressing the components of the required velocity as polynomials in the position

co-ordinates and time, with empirically determined coefficients.

(2) Approximating the physical situation with a model simple enough to yield an explicit solution for the required velocity vector.

The first method has the disadvantage that the determination of the coefficients may require a considerable amount of computation using a large computer and that the values of the coefficients may be rather strongly dependent on the vehicle configuration and the nominal trajectory. Small changes in either may require that the whole set of coefficients be redetermined. The success of the second method depends upon finding a suitable approximation, simple enough to yield a solution, but accurate enough to meet the guidance accuracy requirements.

This paper presents a guidance technique based on a sufficiently simple and accurate model. A flight into deep space has two distinct phases:

(1) The escape hyperbola—the path in the vicinity of the Earth for which the Earth's gravitational field acts alone.

(2) The transfer ellipse—the path in deep space for which the Sun's gravitational field, perturbed by the major planets, acts alone.

The assumption made in this paper is that the guidance objective is satisfied if the vehicle is injected into the transfer ellipse with the desired velocity vector (V_∞), found by iteration before the flight. Translated into requirements in the vicinity of the earth this means that the vehicle is guided to an escape hyperbola which has the same energy as the nominal escape hyperbola and whose asymptote is parallel to the asymptote of the nominal hyperbola. Thus, the n-body problem is replaced by a two-body problem with a central force field, for which an explicit solution for the required velocity can be found. It may or may not be necessary to express V_∞ as a function of the burnout time, depending upon the accuracy required, the expected variations in the burnout time and the relative velocity of the target.

To demonstrate the accuracy of the model, a summary of the results obtained on simulated Atlas Able-4 flights to Venus is presented in Table 1. The vehicle was guided in powered flight so that it had the required velocity vector, as defined in this paper, at thrust termination and no corrections were made in free flight. The accuracy is expressed in terms of the "impact vector", defined as the vector from the center of the target body to the asymptote of the approach path. This definition is used to avoid the "focusing" effect of the gravitational field of the target body.

Table 1

Case number	Miss at Venus* (nautical miles)
1	328·2
2	1107·9
3	1278·7
4	1688·5
5	1635·9

* Change in impact vector.

Case 1: Thrust of all engines 1 per cent high.

Case 2: Thrust of all engines 1 per cent low.

Case 3: Thrust of all engines 3 per cent high.

Case 4: Pitch gyro drifting at 60 deg/hr in powered flight.

Case 5: Yaw gyro drifting at 60 deg/hr in powered flight.

2. DERIVATION OF THE VELOCITY REQUIRED AT BURNOUT

The three components chosen to specify the required velocity vector at burnout are the magnitude, the radial component and a yaw (lateral) component.

2.1 Magnitude of Velocity Required (V_r)

In the two-body problem, when the mass, M, of one body is very much larger than the mass, m, of the other body, one has, from the conservation of energy, that the relative velocity of the smaller body is given by

$$V^2 = GM\left(\frac{2}{r} + \frac{2E}{GMm}\right) \qquad (1)$$

where r is the distance between the bodies, E is relative energy of the small body, and G is the gravitational constant. From equation (1)

$$V_\infty^2 = V^2(r)\,|_{r\to\infty} = \frac{2E}{m} \tag{2}$$

so

$$V_t(r) = \left[GM\left(\frac{2}{r} + \frac{V_\infty^2}{GM}\right)\right]^{\frac{1}{2}} \tag{3}$$

If the engine is cut off when the velocity equals V_t, the energy of the escape hyperbola, and therefore the magnitude of V_∞ will be correct.

2.2 Yaw Component

The free flight path will lie in a plane defined by $\bar{r}$, the vector from the center of the earth to the vehicle, and $\bar{V}$, the velocity at thrust termination. If the velocity vector at infinity is to be parallel to $\bar{V}_\infty$, then this plane must also contain $\bar{V}_\infty$. Thus $\dot{\epsilon}_y$, the component of $\bar{V}$ normal to the plane defined by $\bar{r}$ and $\bar{V}_\infty$, must be zero. Referring to Fig. 2, $\dot{\epsilon}_y$ is computed from

$$\bar{u}_y = \frac{\bar{r}X\bar{V}_\infty}{|\bar{r}X\bar{V}_\infty|} \tag{4}$$

and

$$\dot{\epsilon}_y = \bar{V}.\bar{u}_y \tag{5}$$

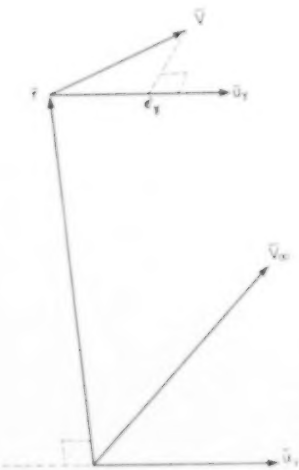


Fig. 2.

2.3 Radial Velocity ($\dot{r}_r$)

If $V = V_t$ and $\dot{\epsilon}_y = 0$, a solution can be found for the radial velocity required, $\dot{r}_{rt}$, for a given $\bar{V}_\infty$. The escape path will be along a hyperbola defined by

$$r = \frac{a(1-e^2)}{1 + e\cos\phi} \tag{6}$$

where

$$a = -\frac{GMm}{2E} \tag{7}$$

and e is the eccentricity (see Fig. 3). From equation (6) and the conservation of angular momentum:

$$(r\dot{r})^2 = GMa\left[e^2 - \left(1 - \frac{r}{a}\right)^2\right] \tag{8}$$

Since a, GM and r are all presumed known, $\dot{r}_r$ can be determined from equation (8) if the required e can be found. From Fig. 3

$$\phi = \pi - (\beta + \theta) \tag{9}$$

so

$$\cos\phi = -\cos\beta\cos\theta + \sin\beta\sin\theta \tag{10}$$

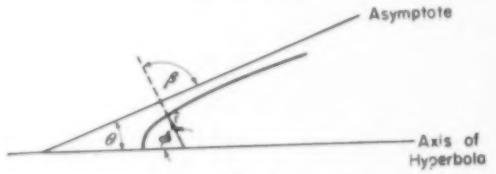


Fig. 3.

Now from equation (6),

$$\cos\phi = \frac{a(1-e^2)}{re} - \frac{1}{e} \tag{11}$$

$$\cos\theta = \cos\left[\pi - \phi\right]_{r\to\infty} = \frac{1}{e} \tag{12}$$

and from Fig. 2,

$$\cos\beta = \frac{\bar{r}.\bar{V}_\infty}{rV_\infty} = x \tag{13}$$

Substitution of equations (11)–(13) into equation (10) yields

$$\frac{a(1-e^2)}{re} - \frac{1}{e} = -\frac{x}{e} + (1-x^2)^{\frac{1}{2}}\left[1 - \left(\frac{1}{e}\right)^2\right]^{\frac{1}{2}} \tag{14}$$

and rearranging equation (14) one obtains

$$(e^2-1) + \frac{r}{a}(1-x^2)^{\frac{1}{2}}(e^2-1)^{\frac{1}{2}} + (1-x)\frac{r}{a} = 0 \tag{15}$$

This is a quadratic in $(e^2-1)^{\frac{1}{2}}$ and can be solved to yield

$$(e^2-1)^{\frac{1}{2}} = -\frac{r}{2a}(1-x^2)^{\frac{1}{2}} \pm \pm\left[\frac{r^2}{4a^2}(1-x^2) - \frac{r}{a}(1-x)\right]^{\frac{1}{2}} \quad (16)$$

Since $(e^2-1)^{\frac{1}{2}}$ is $e\sin\theta$, which is positive because $\theta < \pi/2$, the positive sign is correct. Substituting

$$r' = \frac{r}{a} \quad \text{and} \quad w = 1 + x$$

and squaring equation (16) one obtains

$$e^2 - 1 = \frac{1}{2}r'(2-w)\{r'w - 2 - \sqrt{[r'w(r'w-4)]}\} \quad (17)$$

The required radial velocity can now be found by substituting e^2-1 from equation (17) into equation (8). If the vehicle is to be guided to an ellipse which passes through a fixed point in space with a fixed energy (as might be the case for a flight to the Moon), a similar derivation yields

$$e^2 - 1 = \frac{\frac{1}{2}r'(2-w)\left\{wr' - 2\left(1+2\frac{r}{r_T}\right) - \left[r'w\right]^{\frac{1}{2}}\left[wr' - 4\left(1+\frac{r}{r_T}\right) + 8\frac{a}{r_T}\right]^{\frac{1}{2}}\right\}}{\left(1+\frac{a}{r_T}r'\right)^2 - 2wr'\frac{a}{r_T}} \quad (18)$$

where

$$x = \frac{\bar{r}\cdot\bar{r}_T}{rr_T} \quad (19)$$

and $\bar{r}_T$ is the vector to the target point. Note that equation (18) reduces to equation (17) as $|\bar{r}_T|$ approaches infinity.

3. DERIVATION OF THE STEERING ERROR SIGNALS

In order to specify the error signals it is necessary to select some particular method of steering. The method described here is to guide to a constant thrust attitude such that the velocity vector will be correct at burnout, but not before. A constant attitude trajectory has the following advantages:

(1) The angle through which the missile is turned is smallest so that the yaw component of the radar look angle is smallest.

(2) In yaw it is the optimum for fuel economy.

(3) In pitch, it approximates the optimum for fuel economy.

(4) It leads to a simple mechanization.

3.1 Yaw Attitude Error

Suppose that at the initiation of guidance the missile has some yaw velocity error, $\dot{\epsilon}_y$, and let ψ_y be the angle in the yaw plane between the thrust acceleration and the desired free flight plane, as shown in Fig. 4. The objective is to find a constant ψ_y such that $\dot{\epsilon}_y$ will be zero at burnout. The final yaw velocity error, $\dot{\epsilon}_{yf}$, is given by

$$\dot{\epsilon}_{yf} = \dot{\epsilon}_y + \int_{t_o}^{t_{co}} \ddot{\epsilon}_y \, dt = \dot{\epsilon}_y + \sin\psi_y \int_{t_o}^{t_{co}} a_t(t)\, dt \quad (20)$$

It will be shown in the appendix that

$$\int_{t_o}^{t_{co}} a_t \, dt \cong V_g = V_r - V \quad (21)$$

with V_r calculated from equation (3). So equation (20) can be written

$$\dot{\epsilon}_{yf} = \dot{\epsilon}_y + V_g \sin\psi_y \quad (22)$$

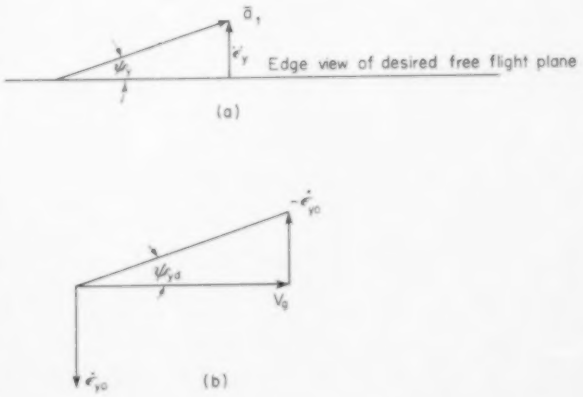


Fig. 4. (a) Illustration of equation (18); (b) illustration of equation (22).

If ψ_y changes to $\psi_y + \Delta\psi_y$, the change in $\dot{\epsilon}_{yf}$ is

$$\Delta\dot{\epsilon}_{yf} = \frac{\partial\dot{\epsilon}_{yf}}{\partial\psi}\Delta\psi_y = V_g \cos\psi_y \Delta\psi_y \qquad (23)$$

Equating $\dot{\epsilon}_{yf} + \Delta\dot{\epsilon}_{yf}$ to zero and solving for $\Delta\psi_y$ yields

$$\Delta\psi_y = -\frac{\dot{\epsilon}_{yf}}{V_g \cos\psi_y} \qquad (24)$$

From Fig. 3, $\psi \cong \bar{\epsilon}_y/a_t$ and $\cos\psi_y \cong 1$ if ψ_y is small. Substituting this into equation (20) and then equation (20) into equation (22) yields

$$\Delta\psi_y = -\frac{\dot{\epsilon}_{yf}}{V_g} = -\left(\frac{\dot{\epsilon}_y}{V_g} + \frac{\bar{\epsilon}_y}{a_t}\right) \qquad (25)$$

$\Delta\psi_y$ is then an attitude error which is to be made zero with a closed loop control system.

3.2 Pitch Attitude Error

A pitch attitude error similar to ψ_y can be determined, but the calculation is more involved because gravity contributes to the radial acceleration. The final radial velocity error is defined by

$$\dot{\epsilon}_{pf} = \dot{r}_r - \dot{r}_f \qquad (26)$$

where $\dot{r}_r$ is the required radial velocity at the predicted burnout position, computed from equations (8) and (17); and $\dot{r}_f$ is the predicted radial velocity at burnout if the missile continues to thrust at its present attitude. If ψ_p is changed by $\Delta\psi_p$, then

$$\Delta\dot{\epsilon}_{pf} = \frac{\partial\dot{\epsilon}_{pf}}{\partial\psi_p}\Delta\psi_p \qquad (27)$$

Since the objective is to make $\dot{\epsilon}_{pf}$ zero, as in yaw, we equate $\dot{\epsilon}_{pf} + \Delta\dot{\epsilon}_{pf}$ to zero and solve for $\Delta\psi_p$, the pitch attitude error.

If $\Delta\psi_p$ is to be progressively reduced to zero by a closed loop control system, then an inaccurate estimate of $\partial\dot{\epsilon}_{pf}/\partial\psi_p$ will be satisfactory. The partial derivative appears in the loop as a gain factor, and the loop can be designed to be insensitive to small changes in gain. From equation (24)

$$\frac{\partial\dot{\epsilon}_{pf}}{\partial\psi_p} \cong \frac{\partial\dot{r}_f}{\partial\psi_p} \qquad (28)$$

and if ψ_p is the angle between the thrust vector and the local horizontal

$$\frac{\partial\dot{r}_f}{\partial\psi_p} \cong V_g \cos\psi_p \qquad (29)$$

then, from equations (25) and (27)

$$\Delta\psi_p = \frac{\dot{\epsilon}_{rf}}{V_g \cos\psi_p} \qquad (30)$$

Note that $\dot{\epsilon}_{pf}$ is the velocity error the missile will have at burnout if it continues to thrust at its present attitude.

3.3 Reduction of Attitude Errors

The design of a closed loop control system to null the attitude errors in the presence of noise gyro drifts and other disturbances is required. This is a separate problem which depends upon the specific system and is not considered here.

4. CONCLUSION

Steering a rocket in powered flight to a set of end conditions which have been determined within the guidance computer during flight is highly desirable. This can be done by defining a desired velocity vector which is a function of the cutoff position and time, and steering the rocket to have this velocity vector. It is economical of fuel to control only the velocity, rather than velocity and position simultaneously.

The required velocity cannot be determined analytically because it involves the solution of an n-body problem. For the guidance of a vehicle whose powered flight phase occurs in the vicinity of the earth, the n-body physical situation can be approximated accurately by a two-body model if the original objective of passing through a point in space is replaced by the objective of attaining a desired escape velocity vector. The accuracy of this approximation has been examined for a flight to Venus and was found to result in errors of less than 2000 miles at Venus for fairly large vehicle perturbations. Errors of this magnitude are smaller than the present uncertainty in the location of Venus. There is no reason to expect significantly larger errors for flights to other targets in the solar system.

The vehicle can be guided to the required velocity vector by a nearly optimum constant attitude thrust profile by relating velocity errors to attitude errors and causing these attitude errors to be zero. The design of a closed loop control system to null the attitude errors has not beeen discussed since this depends on the sources of error and the constraints of the subsystems within the loop.

The guidance technique presented here is not restricted to any particular vehicle except that the major portion of the powered flight must occur in the vicinity of the Earth, and it is applicable to targets anywhere in deep space.

APPENDIX

To show that

$$V_g \cong \int_{t_n}^{t_{co}} a_t \, dt$$

from equation (3)

$$V_R^2 = GM\left[\frac{2}{r} + \frac{E}{GMm}\right]$$

and

$$V_g = V_R - V$$

so

$$\dot{V}_g = \dot{V}_R - \dot{V}$$

$$= -\frac{GM}{r^2}\frac{\dot{r}}{V_r} - [a_t \cos\theta - g\cos\beta]$$

Fig. 5.

with θ and β as defined in Fig. 5. From this figure

$$\cos\beta = \frac{\dot{r}}{V}$$

so

$$\dot{V}_g = g\left(\frac{\dot{r}}{V} - \frac{\dot{r}}{V_r}\right) - a_t \cos\theta$$

or

$$\dot{V}_g + a_t \cos\theta = \frac{g\dot{r}V_g}{VV_r}$$

In the expression on the right, for the Able-4 flight

$$g < 32{\cdot}2$$

$$\dot{r} < 10^4$$

$$V_r > 35 \times 10^3$$

and V_g is monotonically decreasing while V is monotonically increasing; so that

$$\int_{t_n}^{t_{co}} \frac{g\dot{r}V_g}{VV_r}\,dt > \left[\frac{32{\cdot}2 \times 10^4 V_g(t_n)}{35 \times 10^3 V(t_n)}\right](t_{co} - t_n)$$

At the end of guidance in Able-4 (Atlas boosted)

$$V_g \cong 6{\cdot}3 \times 10^3$$

$$V \cong 30{\cdot}5 \times 10^3$$

$$t_{co} - t_n \cong 37$$

so

$$\int_{t_n}^{t_{co}} \dot{V}_g\,dt + \int_{t_n}^{t_{co}} a_t \cos\theta\,dt <$$

$$< \frac{32{\cdot}2 \times 10^4 \times 6{\cdot}3 \times 10^3 \times 37}{35 \times 10^3 \times 30{\cdot}5 \times 103}$$

and, since

$$V_g = -\int_{t_n}^{t_{co}} \dot{V}_g\,dt - V_g + \int_{t_n}^{t_{co}} a_t \cos\theta\,dt < 70{\cdot}3$$

Throughout guidance in Able-4, $\cos\theta$ is about 0·99 and nearly constant; so

$$0{\cdot}99\int_{t_n}^{t_{co}} a_t\,dt - V_g < 70{\cdot}3$$

So V_g is within 2 per cent of $\int_{t_n}^{t_{co}} a_t\,dt$ from the end of guidance to burnout (throughout third stage).

GUIDANCE AND CONTROL OF UNMANNED SOFT LANDINGS ON THE MOON

SIDNEY DARLINGTON
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

Abstract—A guidance and control scheme is described, which may be appropriate for early unmanned soft landings on the Moon. The principal components are: a precision launch phase guidance system; a preset direction and attitude control; main and vernier retrorockets of the solid fuel, fixed impulse type; a range only, pencil beam radar, nonsteerable relative to the air frame; and means for firing the retrorockets at preset radar ranges to the Moon's surface. No means are included for measuring the errors in the burning of the main retrorocket, and yet the vernier is so used that the effects of the errors are largely suppressed.

INTRODUCTION

It is safe to assume that there will be many unmanned soft landings on the Moon, before manned landings are attempted. The first vehicles will carry relatively few instruments, and these will be rugged enough for fairly severe landing shocks, so that the requirements on the landing velocity may be relatively lenient. This paper describes how surprisingly simple the guidance and control equipment actually can be, on board the first soft landing lunar vehicles, and explains guidance principles which make the simplicity possible. Some of the principles may be valid also for the ultimate manned landings.

For definiteness, a specific scheme is outlined, which should be feasible in the near future, with guidance and control components such as now exist. The description assumes that the location of the landing, on the Moon, need not be known to better than 20 or 30 miles, and that landing velocities up to 100 ft/sec are tolerable, provided velocities perpendicular to the Moon's surface are substantially smaller. The landing is to be planned for one of the large generally flat regions (large enough to accommodate the 20 or 30 mile deviations). The vehicle will approach the moon at perhaps 10,000 ft/sec. Reducing 10,000 ft/sec to 100 ft/sec reduces the kinetic energy by a factor of 10,000. Even so, a vehicle landing at 100 ft/sec will need substantial shock absorbers, as well as a rugged pay load. Accuracy requirements on guidance and control components can easily be adjusted for other velocity limits, if desired, but they may not be easily realized if the velocity limits are made much tighter. (Substantially softer landings may require a more complicated terminal guidance system, which reduces the accuracy requirements on the individual components).

A SPECIFIC GUIDANCE AND CONTROL SCHEME

Launch conditions are so controlled that terminal adjustment of the landing location is unnecessary (within the 20 or 30 mile accuracy). Then the terminal guidance and control is used solely to reduce the approach velocity to a tolerable landing velocity. The approach velocity is reduced to landing velocity by means of retrorockets (since the lunar atmosphere is insufficient for aerodynamic braking). The terminal guidance and control must ensure that the retrorockets do, in fact, almost cancel the approach velocity. Furthermore, the cancellation must be accomplished quite close to the Moon's surface. (If the vehicle is stopped at an altitude of 100 ft, and allowed to fall to the surface, it will strike at about 30 ft/sec.)

No landing phase velocities are measured within the vehicle. This is possible because the two velocity vectors illustrated in Fig. 1 can be predicted in advance to sufficient accuracy. One is the vector approach velocity, or impact

velocity in the absence of retrorockets; the other is the net vector change in velocity produced by the retrorockets, at the time the vehicle reaches the Moon's surface. We shall consider the two velocity vectors in turn.

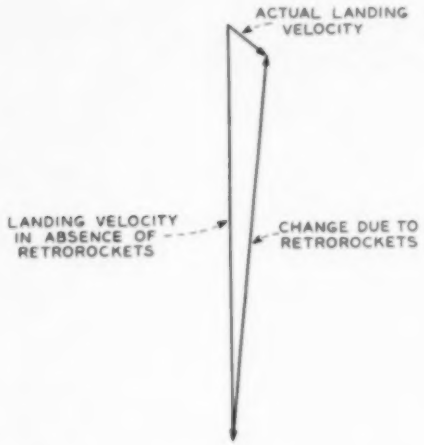


Fig. 1. Vector velocities to be predicted.

The vehicle is launched into a ballistic trajectory, near the Earth. It is then in free flight until the retrorockets are fired, near the Moon. (There is no mid-course guidance.) The vector approach velocity, near the Moon, is to be controlled within the required limits by precise control of the vector initial velocity. Precision (launch phase) guidance systems now in existence may be accurate enough for this. They cannot now be applied to all stages of lunar probe launching rockets, for the rocket-borne parts of the present precision guidance systems are too large for the small final stages of the present lunar probes. However, the precision guidance should become applicable in the near future, as guidance packages become smaller and final stages larger.

To evaluate accuracy requirements, one must relate vector errors in the initial velocity, near the Earth, to resulting vector errors in the approach velocity, near the Moon. In Fig. 2, a vector error dv, in the approach velocity v, is divided into parallel and crosswise components, dv_y and dv_x. The parallel component, dv_y, depends almost entirely on the error, dv_1, in the initial *speed*. Roughly:

$$\frac{dv_y}{dv_1} = 3{\cdot}5 \text{ ft/sec per foot per second.} \quad (1)$$

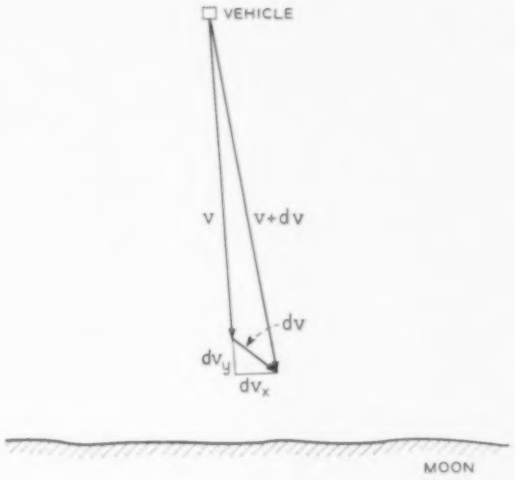


Fig. 2. Errors in the vector approach velocity.

The crosswise component, dv_x, depends on the error, dE_1, in the initial velocity *direction*, as well as on the error in speed. Very roughly, with orbits of the expected sort:

$$\frac{dv_x}{dv_1} = 20 \text{ ft/sec per foot per second*}$$

$$\frac{dv_x}{dE_1} = 500 \text{ ft/sec per angular mil}\dagger \quad (2)$$

If the initial velocity is controlled to 1 ft/sec in speed, and to 0·1 angular mil in direction, dv_y will be restricted to about 3·5 ft/sec, and dv_x to perhaps 60 ft/sec.‡ Horizontal components of these errors contribute directly to the horizontal component of the landing velocity. Vertical components contribute only slightly to the vertical component of the landing velocity,

* This assumes launching from Cape Canaveral, at a favorable time of the month, and at a velocity chosen to minimize velocity sensitivities.

† This refers specifically to *elevation* errors in the initial velocity direction. Sensitivities to *azimuth* errors are substantially smaller.

‡ Allowing for statistical independence of dv_1 and dE_1.

for a reason which will be explained in connection with the retrorockets.

In connection with equation (2), the following relation is of interest: The error coefficients in equation (2) depend almost entirely on the way in which the Moon's gravity bends the orbit of an approaching vehicle. The effect is illustrated in Fig. 3. Deviations in the initial velocity shift the approach orbit almost parallel to itself, with very little change in direction at long distances from the Moon. Then the change in direction at impact, due to the focusing effect of the Moon's gravity, is very closely related to the change in the location of the impact point on the Moon's surface. Under the expected conditions, the relation is as follows: If $d\rho$ is the shift in impact point, in great circle degrees along the Moon's surface, the change in orbit direction is roughly $\frac{1}{5}$ of $d\rho$. The velocity accuracies cited control the landing location to something like 20 or 30 miles.

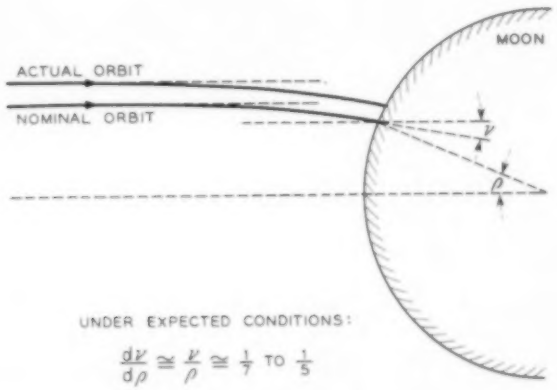


Fig. 3. Curvature of orbits due to the Moon's gravity.

Because of the above relation, the direction reference in the vehicle should refer to space, and not to the local lunar vertical. Deviations in the impact point will shift the direction of the vertical more than the direction of the vehicle's orbit (which coincides with the direction of the velocity vector).

Errors in the initial velocity also cause deviations in the *time* of arrival at the Moon. These are so large that the retrorockets cannot possibly be fired at a preset time. A change in initial velocity of 1 ft/sec will change the time of arrival by something like 100 sec. In 100 sec, the approaching vehicle travels something like 200 miles. As a result, certain minimum observations of the Moon must be made from the approaching vehicle.

In the specific scheme which we are considering, the minimum observations are made with a range only, pencil beam radar, sighted along the direction of the nominal approach velocity. Then the retrorockets are fired on the basis of distance to the Moon, rather than time. The radar does not have to measure any directions, or any rates or velocities. Firing the retrorockets on a distance basis also helps to reduce sensitivities to deviations in the performance of the rockets, as we shall see.

The vector change in velocity due to the retrorockets is designed to cancel the nominal vector approach velocity. If the direction of the thrust is slightly in error, there will be a corresponding crosswise contribution to the residual velocity, comparable to the error dv_x in the approach velocity (Fig. 2). Errors in the direction of the retrothrust will be due primarily to errors in a direction reference within the vehicle. An error of $\frac{1}{4}^\circ$ will contribute roughly 50 ft/sec to the horizontal component of the landing velocity.

The direction reference is preset at launch, so that no direction measurements (star tracking,

lunar horizon tracking, etc.) need be made from the vehicle. The initial setting must be retained through some 2 days of weightless flight (as well as through the launching phase). A gyroscopically controlled stable platform may be the preferred embodiment. A stable platform must be accompanied with an attitude control system or autopilot, so that the vehicle itself can be oriented with the reference direction as it approaches the Moon. Then the retrorockets are rigidly attached to the vehicle (unless they are to be slightly deflected during burning to improve attitude stability).

A common alternative to a stable platform is to spin the whole vehicle, which then becomes its own gyroscope. For this, however, the axis of spin must be the (preset) axis of the retrothrust, which may not coincide with the axis of the last stage of the launching rocket. There is also the problem of stopping the spin at the instant of landing.

The retrorockets are solid fuel units, burned to the end of their fuel, without metering or control of their thrusts or net velocity impulses. If a single such rocket were used, the requirements on its specific impulse and burning rate would be rather too severe. If an integrating accelerometer and thrust cutoff were used to control the net impulse, nonnominal thrusts would still bring the vehicle to a stop at incorrect altitudes.

Reasonable error tolerances are achieved by using two retrorockets, a large or main unit, followed by a small or vernier unit. (Or there may be two vernier rockets acting as a unit, one on each side of the main rocket, as a means of keeping thrust symmetry.) Remarkably enough, the errors in the specific impulse and burning rate of the main rocket do not have to be measured, in order to achieve substantial correction by means of the vernier. The correction actually occurs with no measurement of errors, provided the rockets are used in a suitable way.

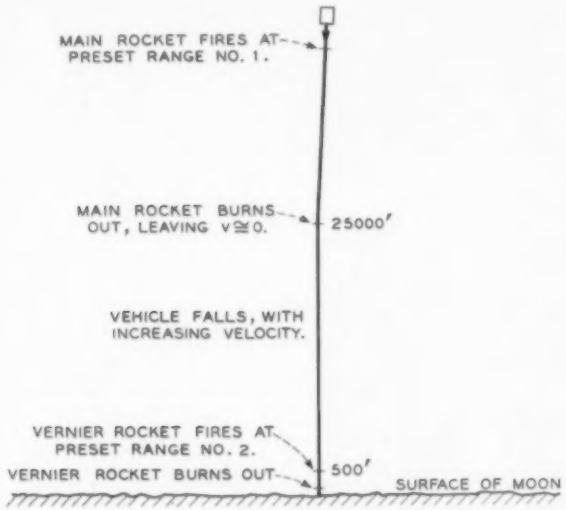

MAIN ROCKET FIRES AT PRESET RANGE NO. 1.
MAIN ROCKET BURNS OUT, LEAVING v≅0.
25000′
VEHICLE FALLS, WITH INCREASING VELOCITY.
VERNIER ROCKET FIRES AT PRESET RANGE NO. 2.
VERNIER ROCKET BURNS OUT
500′
SURFACE OF MOON

Fig. 4. A retrorocket firing sequence.

A specific example is illustrated in Fig. 4. The main rocket is fired by the range only radar, at a first preset range to the Moon, such that the rocket burns out at an altitude of some 25,000 ft. At the instant of burnout, the velocity is nominally zero. For some time thereafter, the vehicle falls toward the Moon's surface, accelerated by the Moon's gravity. At a second, much smaller preset range, the radar fires the vernier retrorocket. The velocity imparted to the vehicle by the Moon's gravity, during the free fall, can be quite accurately predicted. During a 25,000 ft fall it amounts to some 500 ft/sec. The

vernier retrorocket is designed to cancel this velocity, and is fired by the radar at a range which puts the vernier burnout close to the Moon's surface.

The dynamics of the free fall, from the burnout of the main rocket to the firing of the vernier, lead to a substantial suppression of the main rocket's errors. In other terms, the velocity of the vehicle at the firing altitude of the vernier is quite insensitive to deviations in the velocity and altitude of the vehicle at main rocket burnout.

Let the nominal, or design conditions correspond to a free fall through h feet, starting with zero velocity at the burnout of the main rocket and reaching a velocity v_2 at the firing altitude of the vernier. Suppose the actual velocity at main rocket burnout is a vertical velocity v_0, instead of zero. Suppose the actual altitude at main rocket burnout is $h + dh$, instead of h. Let dv_2 be the resulting error in the velocity v_2 after the free fall to the fixed firing altitude of the vernier. Then, for small deviations,

$$dv_2 = \frac{v_0^2}{2v_2} + \frac{g_m dh}{v_2} \tag{3}$$

The coefficient g_m is the acceleration due to the Moon's gravity (some 5 ft/sec per second). The formula is easily derived from the conservation of the combined potential and kinetic energies, during the free fall.

If h is about 25,000 ft, v_2 is about 500 ft/sec. Then a v_0 of 100 ft/sec (instead of a nominal zero) and an error dh of 2000 ft will each change v_2 by about 20 ft/sec. These changes correspond to errors of about 1 per cent in the specific impulse of the main rocket and more than 1 per cent in the burning rate (assuming a burning time of 20 sec).

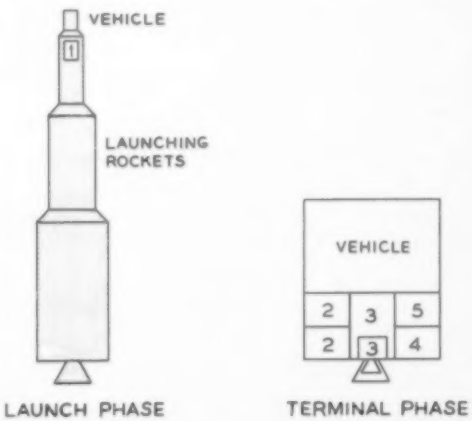
Percentage deviations in the *vernier* performance have a small effect merely because the velocities are relatively small. With a 500 ft/sec vernier, a 2 per cent error in specific impulse changes the landing velocity by about 10 ft/sec. A 2 per cent change in the burning rate changes the altitude at burnout by about 10 ft (assuming a burning time of 2 sec).

The propulsion supplied by the vernier is extra propulsion, not a slice of the minimum propulsion required to stop the vehicle, for the main rocket by itself is sufficient to stop the vehicle. The extra propulsion is the price paid, in this particular scheme, for the lenient accuracy requirements on the main rocket. Different amounts of vernier can be used, if desired, with suitable changes in the altitude of the vehicle at main rocket burnout. The factor by which the main rocket errors are suppressed is proportional to the price paid in terms of extra propulsion.

The scheme works best when the nominal orbit strikes the center of the Moon (as viewed from the approaching vehicle). Then the corresponding vector approach velocity, radar sight line and retrorocket axes are all parallel to the nominal local lunar vertical. The scheme can be adjusted, however, for lines of action which depart considerably from the local vertical, perhaps by as much as 30°. Then the nominal location of the landing can be placed anywhere in a circular area, with a radius of perhaps 500 miles.

CONCLUSIONS

The guidance and control of unmanned soft landings on the Moon can be quite simple, and



1. PRECISION LAUNCH-PHASE GUIDANCE COMPONENTS
2. DIRECTION REFERENCE AND ATTITUDE CONTROL
3. MAIN AND VERNIER RETROROCKETS
4. RANGE ONLY, PENCIL BEAM RADAR
5. RADAR CONTROL OF RETROROCKET FIRING

Fig. 5. Guidance and control components.

can use components of sorts which now exist. Many different arrangements may be possible. The specific scheme described calls for the components listed below, and represented schematically in Fig. 5.

(1) A precision launch phase guidance system controlling the initial velocity vector to 1 ft/sec and 0·1 angular mil.

(2) A direction reference and attitude control system, keeping a preset attitude to $\frac{1}{4}^{\circ}$.

(3) Main and vernier retrorockets of the solid fuel, fixed impulse type, fixed to the airframe (unless deflected for attitude stabilization). Burning rates accurate to 1·5 per cent, specific impulses to 1 per cent for the main rocket and 2 per cent for the vernier.

(4) A range only, pencil beam radar, non-steerable relative to the airframe, directed antiparallel to the retrorockets. Range accuracy perhaps a few hundred feet at 40 miles and 20 ft at 1000 ft.

(5) Means for firing each rocket when the radar range matches a corresponding preset range.

The component accuracies noted should keep landing velocities less than 100 ft/sec, with smaller velocities perpendicular to the Moon's surface. Deviations in the landing location may be 20 or 30 miles.

A FAST TECHNIQUE FOR SATELLITE POSITION DETERMINATION WITH LARGE TRACKING CAMERAS

G. R. MICZAIKA and NANNIELOU H. DIETER
Geophysics Research Directorate, Air Force Cambridge Research Center,
Air Research and Development Command

Abstract—A technique for the fast determination of satellite position from photographs taken with large tracking cameras is discussed. A special "blink-microscope" is described which serves to facilitate the detection of the image. It has provisions for determining the satellite position in an astronomical coordinate system by comparison with a special star atlas.

The basic information required to maintain a continuous surveillance of artificial earth satellites consists of data determining their position in space or as projected on the celestial sphere at a known time. One can distinguish four different techniques for obtaining the desired data: Visual (with or without telescopic equipment), photographic, radio direction-finding and radar. It was determined at System 496*L*, Space Track which has been given the responsibility of developing and operating a space surveillance control center, that none of these techniques alone is capable at the present time of producing data adequate to maintain a satisfactory surveillance. However, each of them can be developed and improved. This report deals with a photographic method of data collecting through the use of large satellite tracking cameras, and the rapid read-out of data in particular.

The requirements to be met if all space vehicles in orbit are to be continuously monitored are of two kinds. One is that the position of the satellite on the celestial sphere should be determined with sufficient accuracy. For our present purposes, surveillance without any specific scientific goals, an accuracy of 1′ of arc and $\frac{1}{10}$ sec of time are considered adequate. Observations of higher accuracy are desirable, of course, and will become more important with improvement of the theory of motion of the vehicles. The second requirement is that of fast producing and handling of the data, on a real-time basis if possible.

Photographic techniques as employed for astronomical work have great advantages for the first of these requirements. Cameras with wide fields are the most suitable instruments for observation of a faint, rapidly moving object if the angular velocity and a crude position are known. In addition the position of the object can be determined on the photographed background of stars to an accuracy of a few seconds of arc or better with such cameras. It is in the second requirement, the fast data-handling, that conventional photographic methods are inadequate. The film must be developed and dried; the satellite image must be found; and the position relative to the background star field must be measured. The entire procedure may very well require an hour or more for a single data point.

The problem of finding a satellite image on a film having a star background is sometimes a difficult one. For example on one frame of a Baker–Nunn satellite tracking camera, which covers $5^\circ \times 30^\circ$ of the sky, there are on the average 200–300 stars of the ninth magnitude or brighter, and in addition between 1000 and 10,000 fainter stars, so that there is about one star for each 2 mm² on the film. When the satellite is as bright as the sixth magnitude, which is the case for a 20-in. spherical satellite at a height of 200 miles, locating the image in the confusion of stellar images is not too serious a problem. However, when the object is as faint as the ninth magnitude (a similar satellite at a height of 1000 miles), finding the image in the

dense star background requires an experienced observer. The relatively rapid angular motion of the satellite with respect to the star background certainly helps in the identification of the satellite image. For a satellite at a height of 200 miles, the image moves 13 mm on the film in 1 sec; however, the image of a satellite at 1000 miles moves only 2 mm in the same time. It is obvious, then, that the identification difficulty increases rapidly with the height of the satellite. The situation is even more difficult for a probe-type vehicle not only because of the increasing faintness of the object, but also because the principal component of its velocity is along the line of sight rather than across it.

An even more serious problem lies in the determination of the position of the satellite image. For this purpose several stars closely surrounding the image must be identified and their positions in the equatorial system (right ascension and declination) found in a suitable star catalog. There are essentially two types of star catalogs of a very different nature. One type of catalog lists individual measurements of star positions made in a uniform system. The so-called fundamental catalogs contain the positions and proper motions of a number of stars well distributed over the sky. Together with a numerical value of the constant of precession, they define the co-ordinate system for a given epoch. The data in the catalog are sufficient to calculate the co-ordinates of each fundamental star for any moment in the past or future. In this fashion the reference system remains defined for any instant.

Fundamental star catalogs are developed from a critical analysis of great numbers of star catalogs which essentially record positions measured with meridian circles or occasionally other techniques. By the analysis the systematic errors which usually are present in the catalogs of individual positional measurements are reduced.

The establishment of fundamental catalogs is one of the major problems of astronomy. At the present time the most outstanding catalogs of this kind are the *Dritter Fundamental-Katalog des Berliner Astronomischen Jahrbuchs* (FK3)[1] and the *General Catalogue* by Boss (GC).[2] Both have extremely small random and systematic errors. The FK3 contains 1535 bright stars and the GC 33,342. Since there are 41,253 square degrees in a sphere it is obvious that these catalogs are not meant for the photographic determination of positions and usually are not suitable because there are in general not enough catalog stars in the near vicinity of the object to be measured. Efforts to provide catalogs of a semi-fundamental nature containing large numbers of fainter stars for reference purposes are under way but these catalogs are not yet available. Instead, one must consult individual listings of star positions, of which a large number is available. Such catalogs frequently cover only isolated zones in the sky and their accuracy differs considerably. A major problem for high accuracy requirements is the lack of knowledge of the stellar motions, since the positions of the stars are given only for the time at which they were observed. If the stars are to be used as reference stars, their positions must be known for the instant at which the plate to be measured was taken. Many years or decades may have gone by between these two epochs. Without knowing the individual proper motion of the stars a computation of the position of the reference stars for the new plate epoch is of much lower accuracy than that of the positions for the original catalog epoch. Selection of a suitable set of reference stars and transformation of their individual co-ordinates into a uniform system is not always possible in a satisfactory way.

When the celestial co-ordinates of the reference stars have been established, the next step is to use them to find the co-ordinates of the satellite. The reference star images should be as close as possible to the satellite image; results are best if their geometric arrangement is of a certain pattern with the satellite image close to the center of the configuration, and if their brightness is similar to that of the satellite. The decision concerning which reference stars to use is one which should be made by an experienced and skilled operator. The location of the star and satellite images must then be measured on the film in an arbitrary system of plate

co-ordinates, and this information reduced to a satellite position in celestial co-ordinates. It is obvious that this whole procedure is not one adapted to rapid calculation of the motion of the satellite. For highly accurate positions there is apparently no way to avoid the lengthy procedures developed and used in positional astronomy. As a matter of fact, there is only a slight reduction in effort if one lowers the accuracy requirements and still follows the conventional procedure. If one can avoid this approach positional data from photographs can be produced in a fraction of the time usually required.

The photographs to be used in the modification of conventional astronomical position determination will be obtained with Baker–Nunn satellite tracking cameras. These cameras were originally developed for the Smithsonian Astrophysical Observatory to track IGY satellites. They are still the most powerful photographic tracking cameras available. The optical system is of the Schmidt type, consisting of a 31-in. spherical mirror, a three-element air-spaced corrector plate with four aspherical surfaces of 22-in. diameter; with a free aperture of 20-in. the system operates at a focal ratio of $f/1{\cdot}0$. The size of the image of a point source is less than 0·001 in. at any point in the 30° field, which is again aspherical. The mounting in its original form is of the altazimuth type; it has, however, three degrees of freedom of motion instead of the conventional two. The fork can rotate around a vertical axis; the gimbal ring can be turned around a horizontal axis in the fork, and the camera body itself is mounted on an axis at 90° to the gimbal axis and allows the camera to be tilted to any angle within 65° of the normal to the gimbal ring. A drive allows the camera body to track with a speed which can be varied between zero and 120′ of arc/sec.

Under System 496*L* a modified version of this tracking camera was developed (Fig. 1). The major modification is in the mounting and drive of the camera. The Space Track cameras can be used in the original fashion, but they can also be converted in a matter of minutes into an instrument with equatorial mounting. The base of the mounting can be tilted in such a fashion that the vertical axis becomes the polar axis, an axis parallel to the axis of the Earth. The gimbal axis which stays horizontal in the altazimuth arrangement becomes the declination axis. A drive rotates the camera around the polar axis at sidereal rate. The intersection of the optical axis with the celestial sphere describes circles parallel to the celestial equator. Desirable as the altazimuth arrangement is for tracking of fast moving satellites at low altitudes,

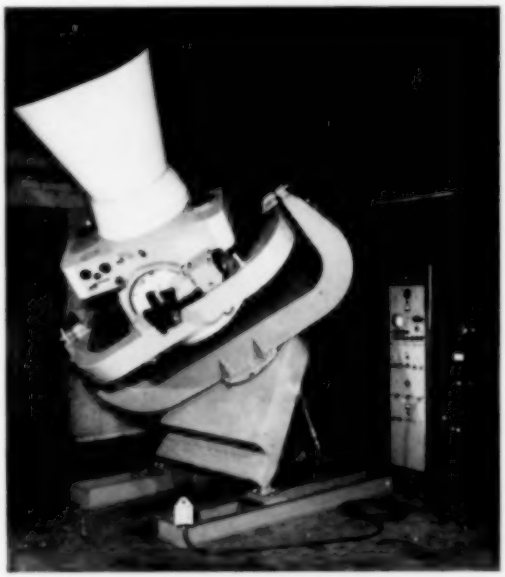


Fig. 1. Baker–Nunn satellite tracking camera in the equatorial arrangement.

an equatorial mounting is preferable when the angular velocity of the vehicle is comparable with the angular velocity of the daily motion of the stars or less than that (satellites in high orbits, space probes, etc.).

The data read-out technique under development does not remove the need for reference stars entirely, but no measurements of relative positions of reference stars on the film or identification of these stars is required, and neither is any appreciable data reduction. There are two major steps in the technique: detection of the vehicle image and co-ordinate read-out. Both steps are accomplished with the help of an

optical device called a "blink miscroscope" (Figs. 2 and 3). Blink microscopes have been used by astronomers for a long time to discover moving objects or stars of varying brightness by comparing two photographs of the same region of the sky taken at different instants. two plates are arranged in the optical train in such a fashion that the star configuration matches in the microscope field. The images appear stationary in the field while the view is switched from one plate to the other except for the image of an object which has moved during

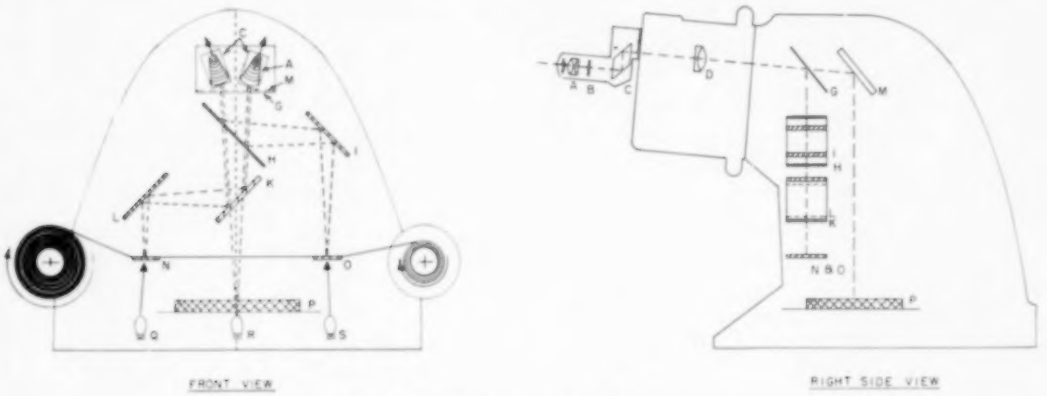


Fig. 2. Blink microscope, optical system.

A	Microscope eyepieces	*H, J, K, L, M*	Mirrors
B	Filter	*N, O*	Film tables
C	Prism	*P*	Atlas table
D	Microscope objective	*Q, R, S*	Light bulbs
G	Semi-transparent mirror		

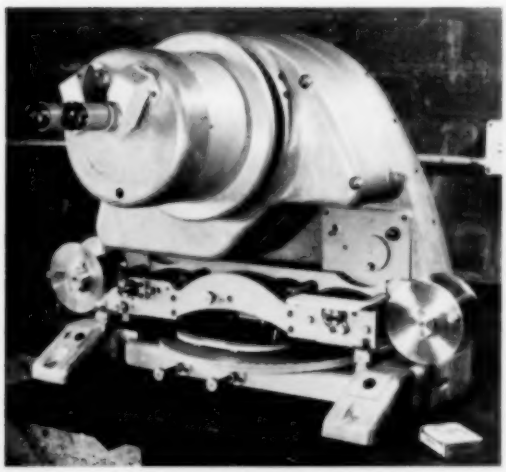


Fig. 3. Blink microscope.

The observer inspects the plates in the field of a microscope of moderate power. A manual or automatic device permits him to switch from one plate to the other in a rapid sequence. The the time between the two exposures. Its image can easily be recognized by its "jumping" in the microscope field. This principle is used for the detection of satellite images by comparing adjacent or more widely spaced frames taken with a Baker–Nunn camera. An especially designed blink microscope is used in which the blinking rate can be varied between zero and five switches per second, and the field of view can be changed by changing the magnification ($3\times$, $6\times$ and $13\frac{1}{2}\times$). The lowest magnification allows to accommodate the full width of the film. Inspection of a pair of frames is a matter of a few minutes, depending somewhat on the density of the star background.

The difference between this blink microscope and the conventional type lies in its third optical channel which serves for the read-out of the co-ordinates of the satellite image. In this channel a properly selected portion of a star map on film can be inserted. This star map has

exactly the same appearance and shape as the satellite films except that its various sections cover $15^\circ \times 15^\circ$ fields of the celestial sphere rather than the $5^\circ \times 30^\circ$ field of the usual frames. Each section of the star atlas carries a grid of co-ordinates in the equatorial system of right ascension and declination. The observer adjusts the satellite frame and the atlas sheet so that the star background in the vicinity of the satellite image appears stationary in the eyepiece while blinking between the atlas section and the satellite frame. Then the satellite co-ordinates are read by a simple interpolation between the co-ordinate grid with the help of an eyepiece reticle.

The Baker–Nunn cameras do not permit photographing fields of $15^\circ \times 15^\circ$ size. Therefore, the standard sheets of the atlas must be obtained by making a mosaic composed of smaller pieces cut from $5^\circ \times 30^\circ$ photographs covering the entire sky. With these remarkably fast cameras all the photographs required can be obtained in a few good nights. At the present time the sixteenth stellar magnitude is considered to be the limiting magnitude of the atlas. On the mosaics the co-ordinate grid will be superimposed. The final star atlas sheets on a film base will then be produced by rephotographing the mosaic with the superimposed co-ordinate grid. The grid will be established in the system of the General Catalogue (for the epoch 1950). The 33,342 stars of the catalogue will serve for the orientation and adjustment of the grid. The complete atlas will consist of about 240 film sheets between glass plates and will cover the entire celestial sphere.

The atlas is primarily intended for the quick data read-out of films taken with large satellite tracking cameras. It will, however, also be a very valuable astronomical atlas which realistically represents the sky as it appears on large field photographs. Its major advantage is the superimposed co-ordinate grid, which permits fast orientation and identification of celestial objects.

REFERENCES

1. A. Kopff, *Veröff. Astron. Rechen-Inst., Berlin*, Nr. 54 (1937).
2. B. Boss, *General Catalogue of 33,342 Stars for the Epoch 1950*, Carnegie Institution of Washington (1937).

THE MILLSTONE RADAR IN SATELLITE AND MISSILE TRACKING*

J. S. ARTHUR, J. C. HENRY, G. H. PETTENGILL and P. B. SEBRING
Massachusetts Institute of Technology, Lincoln Laboratory

Abstract—The tracking radar facility at the Millstone Radar Station of M.I.T. Lincoln Laboratory was designed as a research tool in the study of high-power radar techniques. With the addition of a complex of digital data processing equipment, including the solid state computer CG 24, the Station has found extensive application to the detection and tracking of earth satellites, missiles and space probes. The nature and the performance of both the radar system and the data processing equipment are described in this report and an outline of the data processing programs currently in use in CG 24 is given. The nature and results of several specific operations are discussed.

INTRODUCTION

The Millstone radar was conceived at Lincoln Laboratory in 1956 as a research tool in the study of high-power radar techniques at u.h.f. The size and proposed performance of the radar were considerably in excess of any system prototype up to that time. It was believed necessary, in order to cope with missile detection problems, to meet in a practical way the unknown problems which might be encountered in realizing the design parameters of a system of this size.

The radar went on the air in October 1957 shortly after the launching of the first Soviet satellite. Successful detection of this object was achieved on October 17. In April, 1958, the conical-scan tracking feature was added with the result that Millstone was able to track targets across the sky for a number of minutes, while obtaining continuous position and Doppler measurements. In mid-1958, the CG 24 computer was completed by Lincoln Laboratory and installed at the Millstone Station. With the availability of suitable programs, Millstone became capable of generating ballistic-missile-trajectory impact predictions and of supplying, almost automatically, satellite tracking data. Improvements to both the radar and data processing capabilities are continuing.

* The work reported in this paper was performed by Lincoln Laboratory, a center for research operated by Massachesetts Institute of Technology; this work was supported by the U.S. Air Force.

It is our purpose here to describe and discuss the more interesting portions of the Millstone complex of equipment, the overall performance of the system, and a few of the more interesting experiments which have been performed.

THE RADAR SYSTEM

(a) *Millstone Radar Characteristics*

The basic parameters of the Millstone radar are given in Fig. 1.

(b) *Antenna*

The antenna system shown in Fig. 2 consists of an 84-ft paraboloid on an elevation-azimuth (el-az) mount. R.F. power arrives from the transmitter through the rectangular waveguide, shown along the ground, and through the circular feed horn shown at the focal point of the paraboloid.

The azimuth circule is 90 ft above the ground. The rotating structure on the tower weighs 90 tons. Sixty-h.p. d.c. motors provide slewing rates of the antenna of 4°/sec in azimuth as well as elevation. Accelerations up to 3°/sec² may be obtained.

A turnstile junction, the square box behind the horn, permits the transmission of plane or circularly polarized waves. Control is by adjustment of tuning stubs which are bent up to form a U behind the feed horn. The symmetry of the junction permits simultaneous reception of both the polarization transmitted and that which is orthogonal to it. The conical scan is derived by

Frequency:	440 Mc/s
Antenna:	84′ paraboloid 2° beamwidth 37 dB gain Dual RCVG. polarizations Conical scan tracking
Transmitter:	2 3-cavity klystrons 60 kW ave. power 1·0 MW pk. power 2000 μsec pulse 30 p.p.s. prf Coherent amplifier
Receivers:	200° K eff. temp. 133 d.Bm mds. Variable parameter preamps.
Data processing:	Digitalized output R, $\dot{R}$, α, β, t CG 24 digital computer

Fig. 1. Basic parameters of the Millstone Hill radar.

mechanically rotating the horn which is offset to give a beam squint angle of 0·8° from the axis of the reflector.

Fig. 2. The Millstone Hill radar antenna.

(c) *Transmitter*

Pulsed r.f. power is generated in the transmitter shown in Fig. 3. The klystrons are of the modulating anode type so that high-power pulse transformers are unnecessary. D.C. power is continuously applied, and the electron beam is switched by a pulse applied to the modulating anode.

(d) *Coherent r.f. System*

The same oscillator complex used in synthesizing the transmitted frequency serves to provide the several local oscillator frequencies for the receiving system as shown in Fig. 4. In this manner, shifts in the final 200 kc/s intermediate frequency from the receivers can result only from a frequency (Doppler) shift in the return echo.

Received signals are amplified by parametric amplifiers in both direct and orthogonal polarization channels. These amplifiers afford at least a 3 dB improvement in minimum detectable signal, which corresponds to a doubling of the transmitter power. Intermediate frequencies of 30 Mc/s, 2 Mc/s and 200 kc/s are used. The frequency coherence of return with transmitted signal is utilized at 200 kc/s to derive an unambiguous measurement of the Doppler velocity of the target.

(e) *Filter Banks and Detection Circuitry*

The 200 kc/s i.f. signal from each of the two receiver channels is hard limited and fed to each of two banks of 312 crystal filters which serve in conjunction with associated circuitry to measure Doppler frequency and effect detection of the echo signal. The filters in each bank are spaced at 160-c/s intervals covering the frequency region from 175 kc/s to 225 kc/s which is sufficient to include the expected Doppler range for satellites. The two filter banks are staggered in order to give an effective 80-c/s spacing for cases in which above-threshold echoes are detected in both polarizations. Associated circuitry permits encoding of Doppler to ± 40 c/s for such cases or to ± 80 c/s when a return is received in only one polarization.

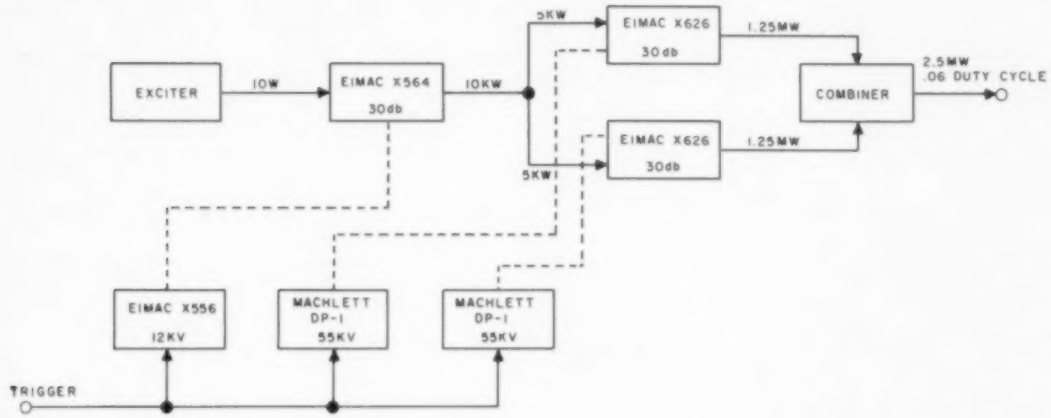


Fig. 3. Block diagram of the Millstone Hill transmitter.

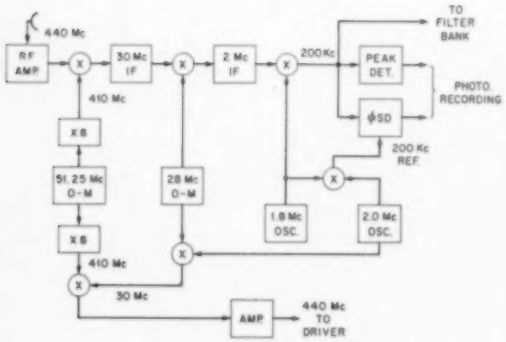


Fig. 4. Block diagram of the Millstone Hill receiver–exciter system.

A filter bank and its associated circuitry is shown in Fig. 5. It will be noted that a pulse, generated by detection of the trailing edge of the received echo, samples the reading in the range counter. The range counter is reset to zero by the trailing edge of the transmitted pulse. The Doppler coding nets code the number of the filter having the greatest response. Further logical circuitry, not shown here, is provided to combine the information from the second filter bank before passing to the remainder of the data processing complex.

(f) *Detection Capability*

Performance of the system is such that single pulse detection can be obtained on a 1-m^2 target at distances in excess of 1000 nautical miles. A plot of the computed detection capability can be seen in Fig. 6.

(g) *Typical Coverage*

A quantitative look at typical Millstone coverage geometries may be of interest in gaining a feeling for the capabilities of such a machine. exhibited a peak cross section of about 20 m² and which has a cross section below 1·0 m² about 30 per cent of the time. The shorter (900 nautical miles) acquisition range is selected in

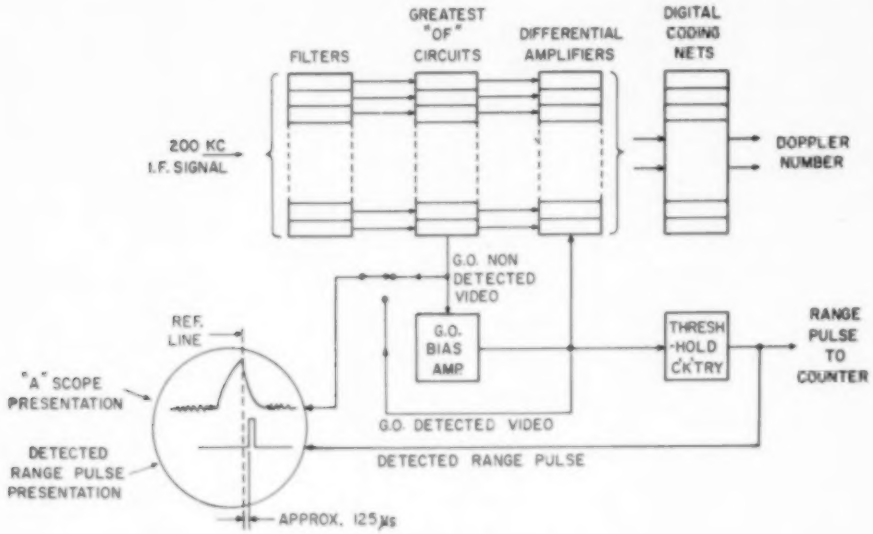


Fig. 5. Doppler filter system which provides Millstone Hill radar with unambiguous Doppler frequency measurements.

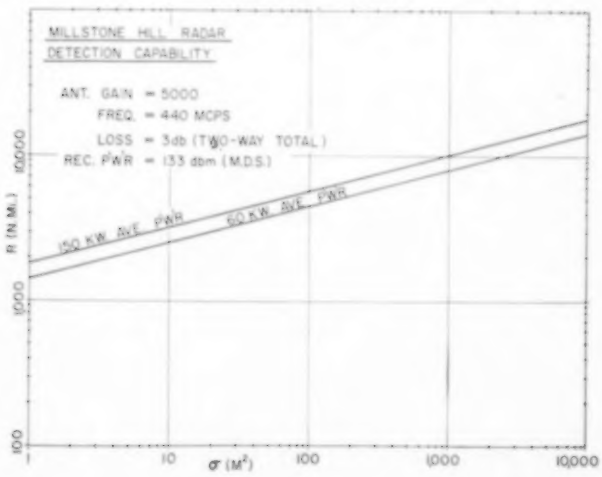


Fig. 6. Theoretical curve of range capability of Millstone Hill radar as a function of target cross section.

Figure 7 shows the plot on a map of a typical track of the instrument package of the Sputnik III (58 Delta II) satellite, an object which has order to increase the probability of acquisition to almost unity.

Figure 8 shows plots of typical missile tracks

as observed at Millstone. In each case, the tracks are limited by horizon at both ends.

(h) *Nature of Echoes*

Echoes of the 2 msec transmitted pulse rising exponentially to 68 per cent of maximum amplitude in 2 msec, then decaying. Echoes from the Moon are distorted and elongated due to varied distances of portions of the Moon's surface. Pulse-to-pulse changes of the later

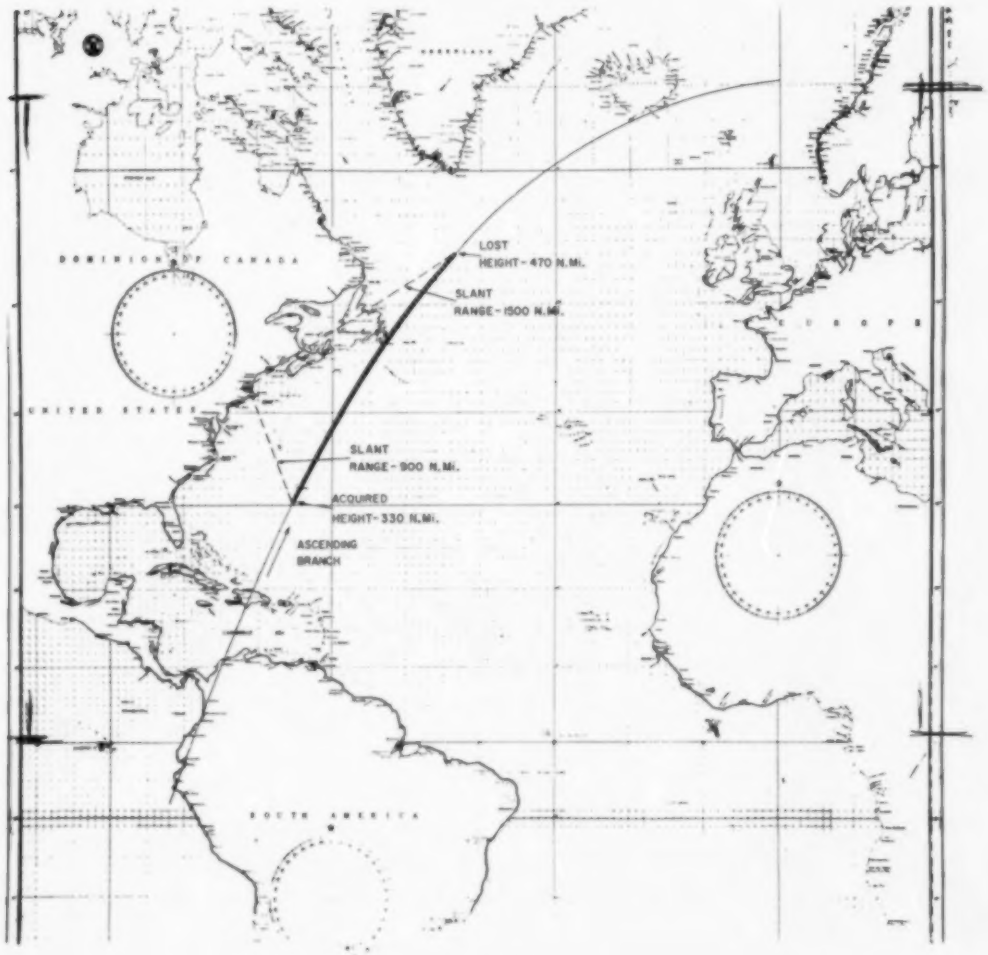


Fig. 7. Millstone Hill coverage on 58 Delta II satellite (Sputnik III).

reflected from the Moon, aurora and a satellite, respectively, are pictured in Fig. 9. The satellite signal, which was obtained from the first Sputnik, is characteristic of the filter response: portions of the echoes arise primarily from lunar libration. At periods of low libration, these changes take place at a correspondingly low rate.

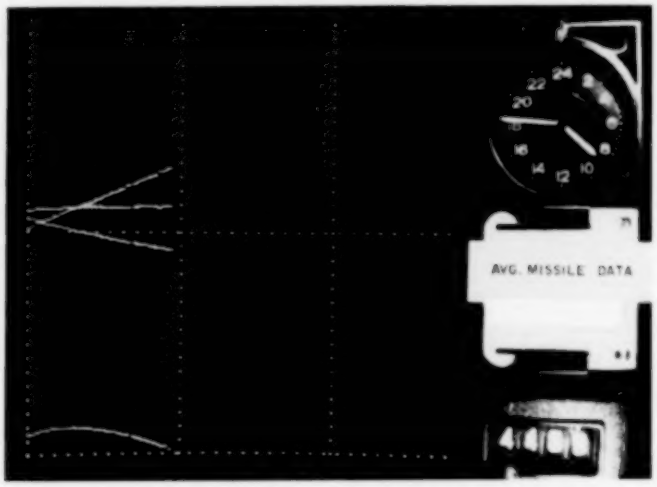


Fig. 8. Computer output display of averaged data plotted against time for a missile track.

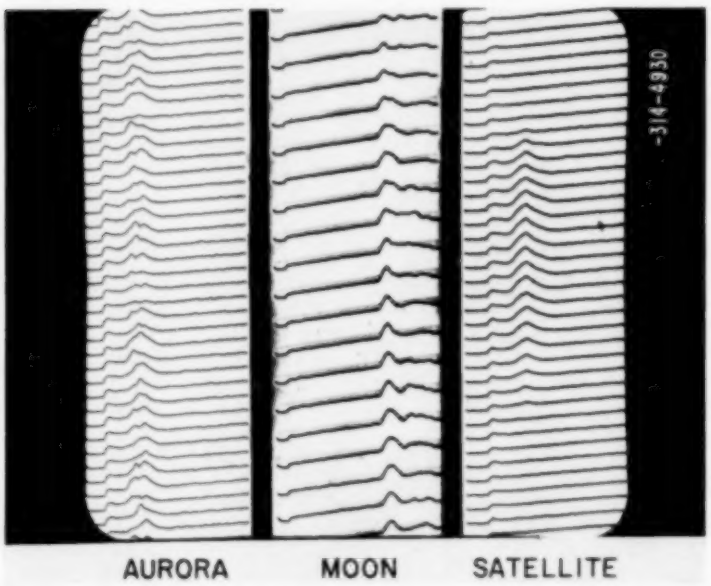


Fig. 9. Amplitude scans of echoes of the Moon, aurora and a satellite as seen by the Millstone Hill radar.

DATA PROCESSING

(a) *General*

Figure 10 is a block diagram of the data processing equipment for the 440 Mc/s tracking system. Echoes from the receiver are detected in either direct or orthogonal Doppler filter bank. When an echo is present, the angles and range are sampled as digital information. The Doppler frequency corresponding to the output from the most strongly responding filter is coded in digital form. All data, including range, range rate, angles and time, are coded in both binary and decimal form. Decimal data are continuously recorded on a high-speed printer shown in the top center of Fig. 10. Binary data are fed to a computer via a buffer. The buffer

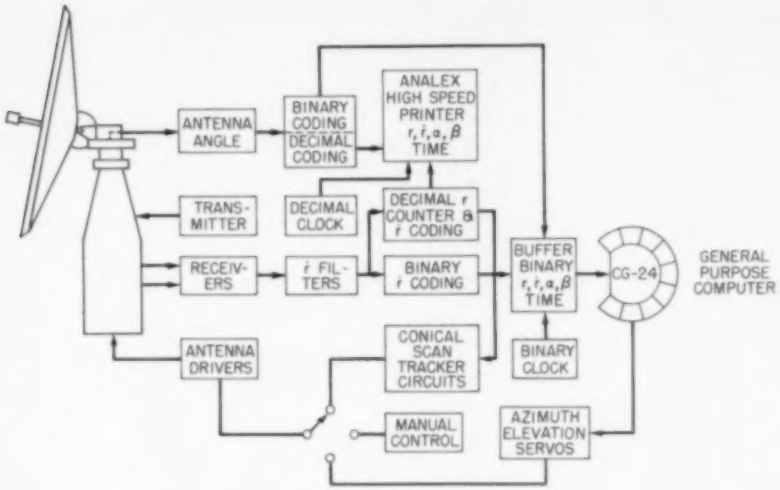


Fig. 10. Block diagram of Millstone Hill radar data processing system.

synchronizes the data to computer timing and provides a two-word buffer store. One output of the computer shown at the bottom of the diagram is a digital servo unit which provides for antenna positioning from the arithmetic section of the computer.

data. Many of the features of the machine are designed to facilitate reliable processing in real time of ballistic trajectory data.

CG 24 devotes some operating time to programs of a non-operational nature. Primarily, however, programming efforts and machine

Measurement	No. binary bits	Least significant bit
Time (t)	24	1/128 sec
Range (r)	12	1 naut. mi.
Range rate ($\dot{r}$)	13	80 c/s ($\approx$80 ft/sec)
Elev. angle (α)	11	0·05°
AZ. angle (β)	13	0·05°

Fig. 11. Word length of binary data sent to CG 24.

A list of the data being sent to the computer is shown in Fig. 11. Column 2 shows the total number of binary bits for each of the measurements of column 1. The third column is the magnitude of the least significant bit handled; the quantizing interval.

(b) *CG 24 Computer*

The computer, which can be programmed to process tracking data in real time, has capabilities as summarized in Fig. 12. Three live registers permit direct storage of real time radar tracking

time are spent on the processing of data for use in prediction, the study of various prediction techniques, the study of the data *per se*, and research into the various operations done by a radar of the type at Millstone. The presence of an integrated computer allows one to simulate and study scanning and acquisition techniques, tracking performance and measurement accuracy. Most error studies of orbit parameter predictions based on radar measurements have assumed Gaussian distributions which, of course, may not be valid. The type of error distribution is directly

measurable, and in some cases may pin-point an unexpected source of bias error.

In the operational mode, CG 24 either stores raw data, i.e. unprocessed binary data, or processes the data before storage. The raw data is used for study of the data, for study of redundant data prediction schemes and for the making of permanent records. Processed data results from a data averaging program (DAP) which subjects the raw data to gating criteria in range, Doppler velocity and return density before averaging. Six-sec intervals are presently being used for averaging as a compromise between linearity of data and the hunting period of the tracker. The file of processed data is always printed out as a record of each track along with photographs of the averaged data plotted against time. It is this averaged data which, in the case of satellite tracks, is made available to users. Computer programs punch teletype tapes directly from memory so that data may be sent semi-automatically in the shortest possible time.

Tracking of missile firings from Cape Canaveral and Wallops Island result in data from which impact predictions are made. For ballistic trajectories several minimum data impact prediction programs are available. ("Minimum data" refers to the fact that the minimum convenient number of measurements are used which are sufficient to determine the orbital elements of the trajectory.) One program frequently used is a semi-automatic prediction routine which requires only that the operator actuate a switch to indicate end-of-track. This program achieves some of the results of redundant data processing by repetitive use of different combinations of minimum data sets and averaging the results. Here it should be noted that normal tracks are from 2 to 10 min in duration, yielding upwards of 50–100 6-sec averaged points.

GENERAL	
Construction:	Solid state
Application:	General-purpose plus real-time control
Timing:	Synchronous, 330 kc/s.
Operation:	Sequential, subject to self-modification
NUMERICAL SYSTEM	
Internal number system:	27-bit binary words, including two parity bits
Single-address instructions	
Fixed-point arithmetic system, programmed floating point subroutine.	
ARITHMETIC UNIT	
Addition time:	24 μsec (incl. memory access)
Mult.-div. time:	84 μsec (incl. memory access)
Square root time:	300 μsec (incl. memory access)
STORAGE SYSTEM	
8192 words, coincident current magnetic cores, 12 μsec cycle time.	
TERMINAL EQUIPMENT	
Flexowriter	
C.R.T. display with numeric generator	
Photoelectric tape reader.	

Fig. 12. CG 24 parameters.

Figure 13 is a block diagram of the program. Averaged data are formed into sets. For prediction equations using three-position measurements, the file is divided into five sets; sets 2 and 4 are not used. Combinations of positions from sets 1, 3 and 5 are used to determine impact points. This normally results in 100 to 200 predictions. All predictions are plotted on the C.R.T. display along with latitude and longitude of the center of this scatter in diagram.

Grid lines are formed through the computed center of the diagram. In addition to coordinates of impact, such data as standard deviation of the two dimensions of the scatter

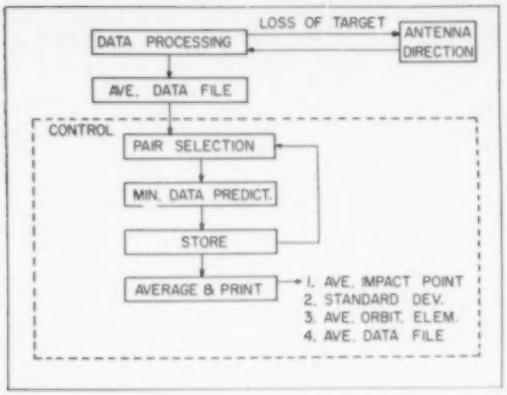


Fig. 13. Block diagram of a semi-automatic prediction routine programmed on GC 24.

diagram, ground range, maximum height, number of points, etc. are printed in the upper right-hand portion of the display. Provisions exist for excluding bad position data points by operator intervention through the Flexowriter. Re-evaluating the impact prediction then may follow.

In summary, averaging occurs during tracking phase, the end-of-track signal is given the computer by an operator when the target disappears or becomes too erratic to track well. Automatic predictions and display then occur giving impact data in about 30 sec after end-of-track.

Figure 14 shows a scatter diagram of the type described above. The total spread here is 100 miles long by 10 miles wide. Averaged data are printed and displayed following prediction. A photograph of a display of averaged data taken on 58 Epsilon, pass No. 3050, is shown in Fig. 15. This is the Explorer IV satellite, a cylinder 80 in. long by 6 in. in diameter. Time is represented along the abscissa beginning with target acquisition. Grid spacing is 2 min. The top curve is azimuth angle (which passes 360° and re-enters the bottom at 0°). Doppler velocity appears as the S-shaped curve, beginning at −20,000 ft/sec and ending at +20,000 ft/sec. The peaked curve is elevation angle, which reaches a maximum of 47° while range appears as a parabolic curve with a minimum value of 400 nautical miles. The target was lost at about 1100 miles. (An analysis of the raw data shows a tumble period of about 12 sec.)

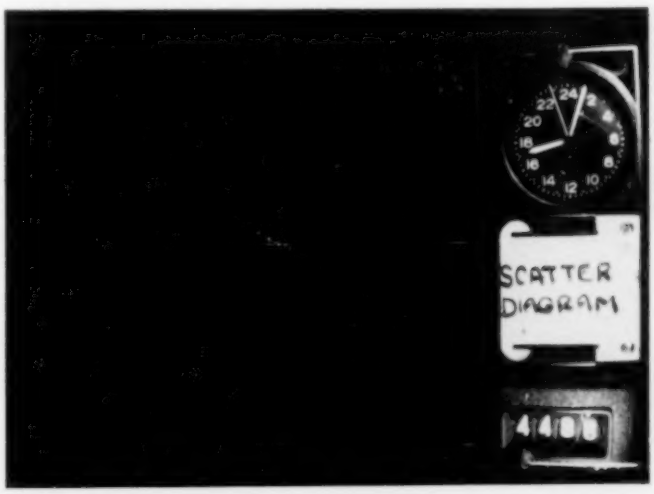


Fig. 14. Scatter diagram of predicted missile impact points determined by repetitive predictions using minimum data equations.

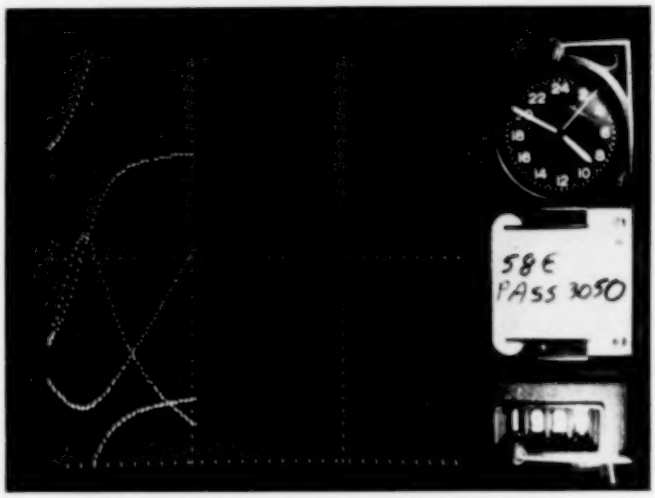


Fig. 15. Output C.R.T. display on CG 24 showing averaged data taken on satellite 58 Epsilon (Explorer IV) plotted against time.

ACTIVITIES

The Millstone facility carries on many activities other than tracking and should not be regarded primarily as a tracking facility. Because of the nature of this discussion, however, the tracking activities are given most emphasis here.

(a) *Satellite Tracking*

Hundreds of tracking operations on satellites have been carried out at Millstone. These targets provide excellent opportunity for system exercise, test of equipment or technique improvements, and for computer program developments. Position data are available to agencies who find them useful. Satellites 57 Beta, 58 Delta I, 58 Delta II, 58 Epsilon, 58 Zeta, 59 Gamma and 59 Delta have been tracked.

Death watches, e.g. the tracking of all visible passes when decay of a satellite is imminent, have been mounted on 57 Alpha I and II, 57 Beta and 58 Delta I to help improve our knowledge of atmospheric drag.

(b) *Missile Tracking*

Most of the benefits derived from satellite tracking hold for missiles as well. In addition, the Lincoln project on re-entry physics is vitally interested in data from certain firings at Wallops Island. Prediction programs are under continual development. The present effort at Millstone involves redundant data prediction following the maximum likelihood techniques as described by Shapiro.

(c) *Other Experiments*

Several space-probe experiments have involved Millstone. Active (skin) tracking is usually done only during the first few minutes of these trajectories as an aid to obtaining injection conditions. Microlock telemetry and angle data are obtained in a passive tracking mode, utilizing the gain of the big paraboloid reflector, but with modified antenna feed structures. On the lunar probe operations in October and November, 1958, active tracking of the second stage past burnout time allowed Millstone to compute position of second stage burn-out. This information was furnished via telephone to the computer at the firing range to aid in the determination of guidance corrections for the final stage. Many other experiments of various types are in process, including high-power component testing, Moon bounce communication, aurora transmission and reflection effects, Venus echo measurements, interferometer studies and meteor echo studies.

(d) *System Improvements*

It was mentioned earlier that system improvements with respect to both equipment and techniques constitute a major activity. Two of the more important hardware developments are the installation of a coded pulse transmitter–receiver developed by Siebert and Kraft, and the installation of a synchronous detection system, designed by General Electric. These devices should permit range and Doppler accuracy improvements, respectively, by approximately one order of magnitude. New parametric amplifiers are installed and under test in the hope of obtaining noise figures of the order of 1 dB.

Digital tape equipment is being installed for computer and site data storage. Amplitude digitizers are also being installed to facilitate the study of cross section.

As an example of system improvement studies, Figs. 16, 17 and 18 depict the preliminary results of error distribution in range measurements. Range is determined by differentiating the output of matched filters (see the waveform on Fig. 5). Each determination is based on single pulse detection and as such is an independent range measurement. The plots are derived from canned raw data for a missile track, hence, many signal amplitudes are represented.

Figure 16 is a distribution of range measurements normalized in range. Errors of more than one type are known to contribute to this distribution. For weak signals, noise causes inflections before true range and is interpreted as a range too short. This accounts for the long tail to the left of the peak. Also, weak signals, disregarding noise, cause the decay slope to be flatter than any fixed system setting and yield some range measurements above true range. This is the shoulder above the peak. Dissimilarities among the filters contribute to the half width of the distribution as do inaccuracies in normalization for the plot. Strong signals have the tendency to produce ranges slightly short of

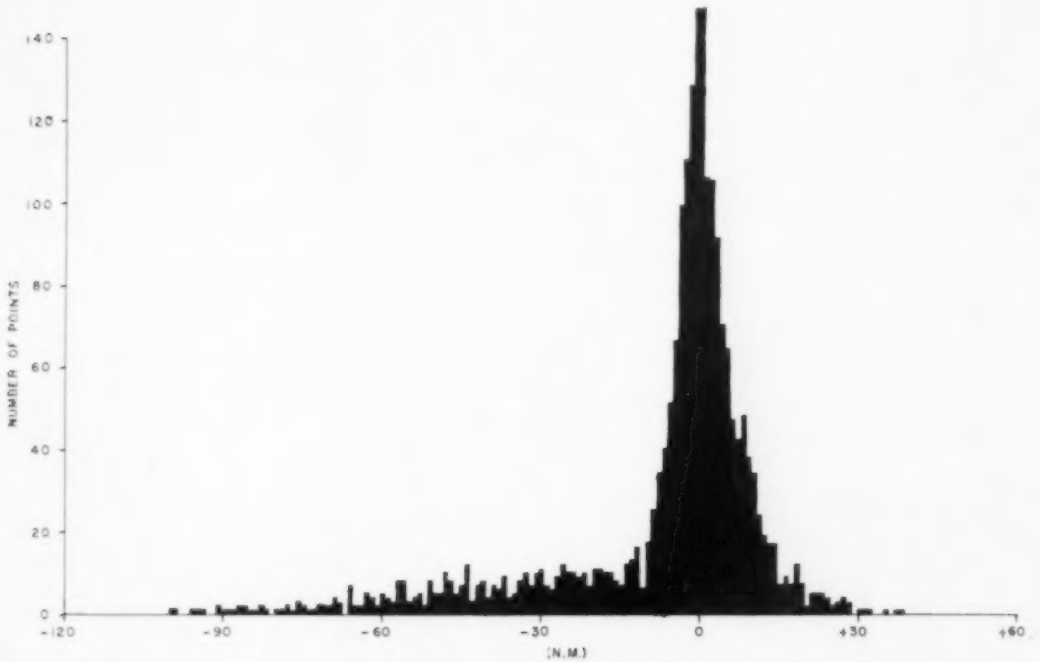


Fig. 16. Unprocessed range error distribution.

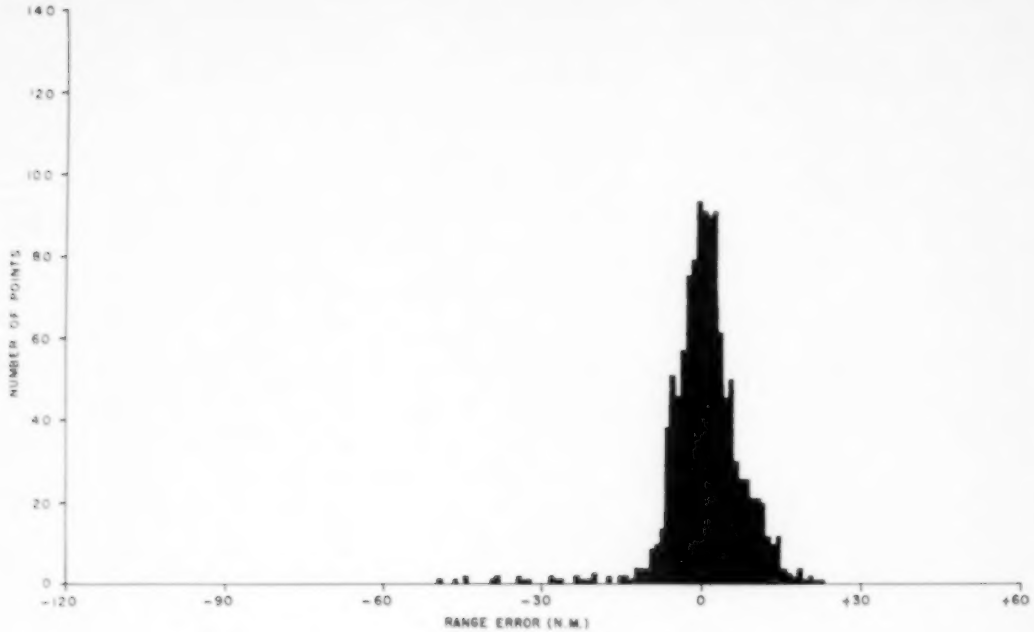


Fig. 17. Range error distribution following data averaging program processing.

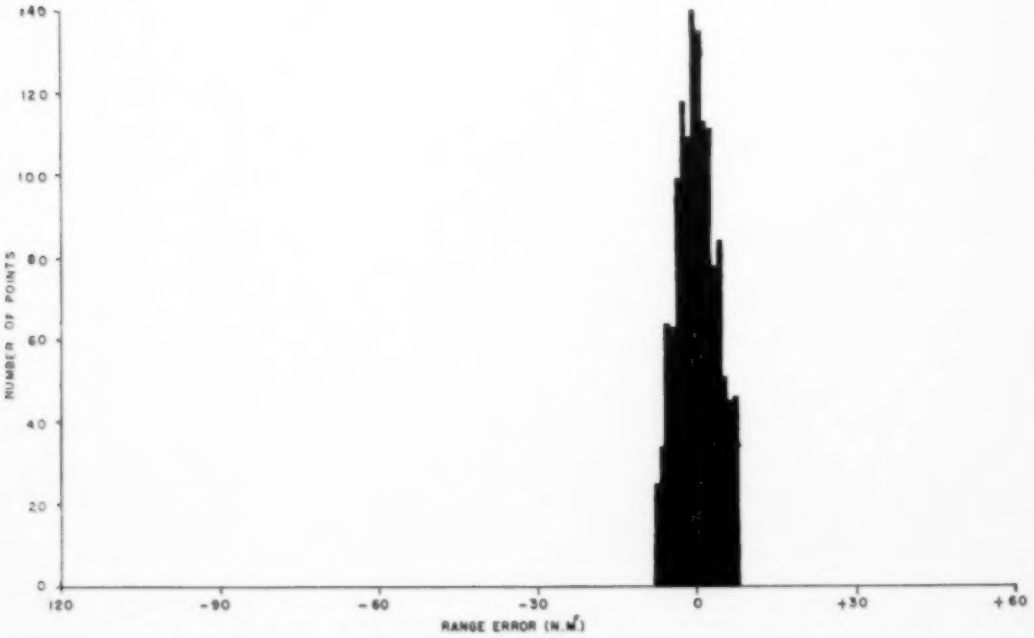


Fig. 18. Range error distribution following processing by a narrow gate covering the time region about signal inflection point.

true range by having a steeper slope than that for which the system is set.

Figure 17 shows the same data after being exposed to the data averaging program. Notice that in addition to throwing away some of the poor ranges, the program decreases the number of good points by about 30 per cent.

Figure 18 is the output from a digital filter in CG 24 which has the effect of making the distribution more symmetric. An analog range gate of narrow width (700 μsec) derived from the conical scan tracking gate duplicates the distribution shown in Fig. 18. Single pulse range determination equipment which shows even narrower and more symmetric error distributions has been developed by A. Galvin of Lincoln Laboratory.

Controlled experiments using known signal–noise ratios are in process to enable better initial adjustment of threshold for given targets, and possible automatic assignment of measurement errors in prediction programs according to measured amplitude.

CONCLUSION

In summary, the Millstone Station has performed successfully as a missile and satellite tracking complex for nearly 2 years. The versatility of the available equipment has permitted useful participation in many experiments involving space probes, satellites and special missile firings. Basic radar accuracies are limited primarily by the use of conical scan tracking and by the use of a long uncoded radar pulse. The latter is being remedied by pulse compression, but the basic angle accuracy will not improve much, for targets with scintillation, unless a monopulse feed is installed.

Acknowledgement—The authors are indebted to the staff of the Millstone Hill facility for their assistance in providing pertinent data with which to compile this paper, and to Arlene Pioli for its final preparation.

A SIMPLIFIED HIGH PRECISION 200 MILLION MILE TRACKING, GUIDANCE AND COMMUNICATION SYSTEM

JAMES W. CROOKS, JR.
Convair (Astronautics) Division of General Dynamics Corporation

Abstract—This paper describes techniques suitable for tracking and guiding missiles to Mars and Venus or beyond. In simplest form, a system can be constructed utilizing these techniques, which requires only two ground stations a few thousand miles apart, linked only by voice or teletype communications and a transponder in the spacecraft, to provide precise three-dimensional tracking information. One dimension is determined by making use of Earth motion, another dimension by the difference of distance from the spacecraft and the ground stations, and the third by direct range measurement.

INTRODUCTION

For the past several years, an Azusa Mark I system has been used at the Atlantic Missile Range with most of the ballistic missiles fired from that range. Well over a hundred missiles carrying Azusa transponders have been tracked by the Azusa Mark I. Now a second system, an Azusa Mark II, is being installed at Cape Canaveral to retire the earlier model Azusa Mark I. It is anticipated that this new system will be the world's most precise radio tracking system for a period of time.

This paper describes a feasible system employing many of the Azusa techniques, but also employing some additional techniques adapting it to the special purpose of precision tracking ranges of spacecraft as far away as the nearer planets.

The intent of this paper is to illustrate the use of a combination of techniques by describing a typical system design employing them. The system design, used for illustration, is called the 2-10-8 System because of the nominal design range of 2×10^8 nautical miles with a low-gain

Fig. 1 Azusa Mark I installed at Cape Canaveral, Florida. Photo taken shortly after system was put into operation.

(14 dB) space vehicle antenna. When used with a high-gain (34 dB) antenna the useful range is extended to 2×10^9 nautical miles and the system becomes the 2-10-9 System.

Fig. 2. Azusa Mark II during setup for tests in San Diego, Spring 1959.

BACKGROUND

In the Azusa systems, range is obtained by measuring the phase delay of modulation signals transmitted from the ground to the transponder which offsets the carrier frequency and retransmits the modulation signals back to the ground.

A special feedback technique is used which employs the same klystron for transmitting as well as for the local oscillator in the transponder. The use of feedback allows extreme precision in range measurement by including variations of delay in the i.f. amplifier in the feedback loop.

In order to provide very sensitive measurement of range rate, a coherent carrier modification was added to the Azusa ground stations and to most Azusa transponders in current use. This modification involves phase-locking the microwave klystron oscillator in the transponder to a received reference frequency. The klystron is locked in phase to within a fraction of a radian of the reference phase at a frequency of 5000 Mc/s by means of a high-gain feedback loop. Since lock must be preserved even under conditions of high vibration, the bandwidth of the receiver providing the reference signal must be several kilocycles to allow feedback loop gain at frequencies up to a few kilocycles. This is necessary to overcome klystron frequency perturbations.

The transponder carrier phase is locked to the carrier signal received from the ground station, offset by a multiple of the highest ranging modulation frequency on that carrier signal. The carrier phase of the transponder signal received back on the ground, when compared with the carrier phase transmitted from the ground station offset by a proper multiple of the modulation frequency, exhibits a one-cycle change of relative phase for a change of missile range of one-half the wavelength of the 5000-Mc/s carrier (approximately 0·1 ft/cycle of phase change). This coherent carrier (Doppler) ranging technique in actual operation provides a measurement of change of tracking range to a

sensitivity of about 0·1 ft even for IRBM's and ICBM's under full acceleration at hundreds of miles of range.

Modifications to increase Azusa tracking ranges to about 50,000 nautical miles by means of bandwidth reduction and correlation techniques are now under consideration. It is not feasible to modify the present Azusa systems to ranges of hundreds of millions of miles. However, many Azusa techniques lend themselves to systems designed for such range.

BASIC 2-10-8 SYSTEM

The system described here is largely based on techniques developed in the Azusa project plus certain new techniques.

The transponder in the 2-10-8 System transmits a fixed frequency microwave carrier signal controlled by a precision crystal oscillator. This carrier signal is modulated by signals produced by dividers in the transponder operating from the same crystal oscillator that controls the carrier frequency. The transponder transmits "clean" signals whether or not it is receiving a signal. The transponder also has the capability of transmitting modulation signals which are a function of the difference in frequency and phase of the signals already described and the signals being received by the transponder. The technique utilized allows a simple narrow-band receiving system.

Figure 3 illustrates the basic principle of the 2-10-8 System. The transponder is transmitting a frequency f_t' in the diagrams and receiving the frequency f_r'. The transmitting frequency is controlled by a multiplier chain multiplying the frequency from a highly stable crystal oscillator. A block diagram of the transponder is shown in Fig. 4. The microwave power amplifier output signal is also used as a local oscillator signal for the microwave mixer in the transponder receiver. A small amount of the transmitted signal feeds through the duplexer filter to the receiver microwave mixer. The difference between the transmitted frequency and the received frequency appears in the i.f. amplifier.

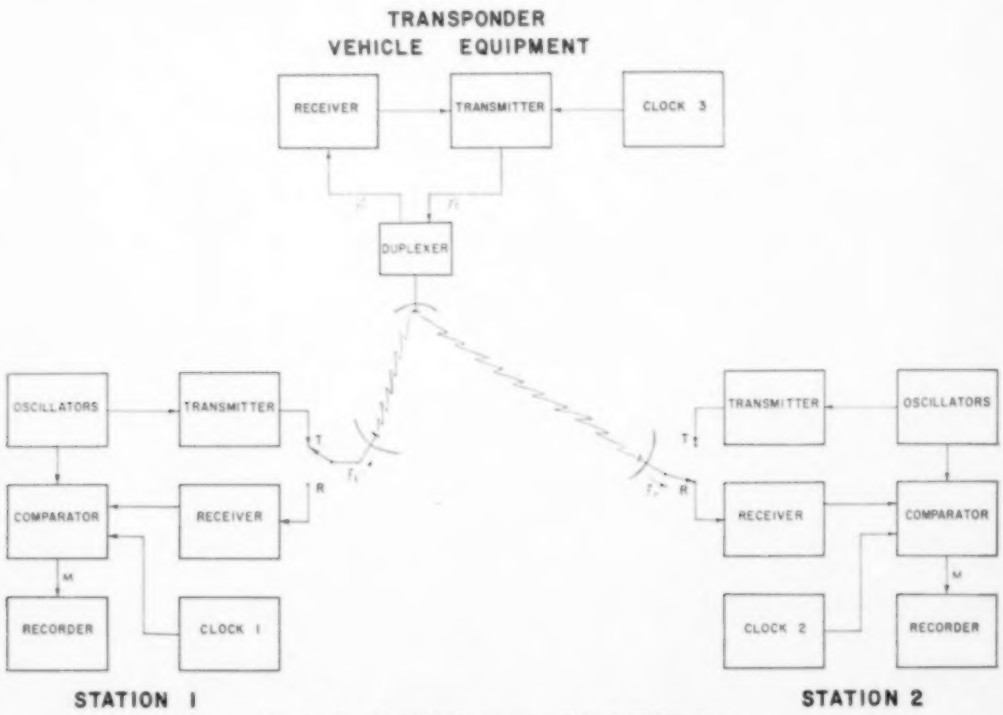


Fig. 3. Simpfilied block diagram of the 2-10-8 System.

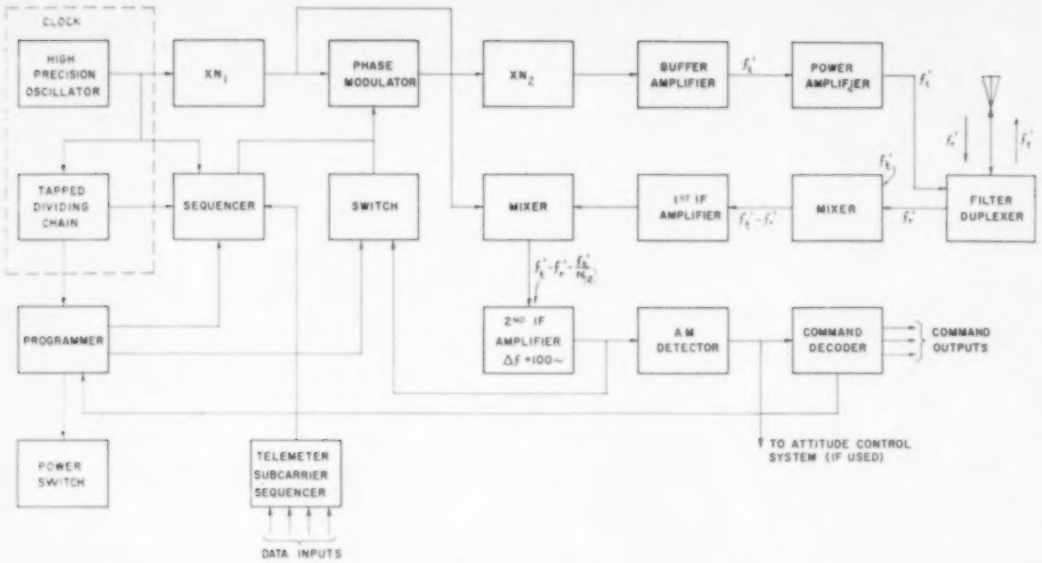


Fig. 4. Block diagram of the 2-10-8 transponder.

TRANSPONDER DOPPLER OPERATION

In order to make the explanation of the principles involved more straightforward, all modulation signals not involved in the coherent carrier (Doppler) operation of the system will be neglected in explaining this portion of the system operation. The additional modulation signals are discussed later.

A signal picked out of the multiplying chain between the crystal oscillator and the microwave power amplifier is used as a second local oscillator signal. It is fed into a second mixer, which in turn produces a very low second intermediate frequency. The second i.f. amplifier has an extremely narrow bandwidth. The output of this amplifier is fed through a program controlled switch to a phase modulator in the multiplying chain controlling the carrier frequency. Operation of the programmer is discussed later. The following discussion applies to the transponder programmed to transpond received signals. The difference in the frequency of the carrier signal being received and the carrier frequency being transmitted (offset by a subharmonic f_t'/N_2) modulates the transponder-transmitted signal. Gain is limiter-controlled so that the peak voltage out of the second i.f. amplifier is held constant when the signal being received is well above the ambient noise in the receiving channel.

GROUND STATION DOPPLER OPERATION

On the ground the received carrier frequency is compared with the reference carrier frequency (f_{t_0}) with ground station offset by a subharmonic (f_{t_0}/N). A technique similar to that used in the transponder is utilized.

The multiplying ratio N in the ground station is made equal to N_2 in the transponder. Fig. 5 is a block diagram of the ground station. The output of the second i.f. amplifier, which is the desired difference frequency, is filtered in the tracking filter (No. 1) and is digitized and recorded on recorder channel T_5.

The frequency of the ground station oscillator (No. 1), which controls frequency f_{t_0} is digitized and recorded on channel T_3. Since the digitizer read pulses are generated by an atomic controlled frequency reference, all digital measurements are referred to this "clock" frequence reference.

The carrier component of the modulation on the transponder signal due to the transponder

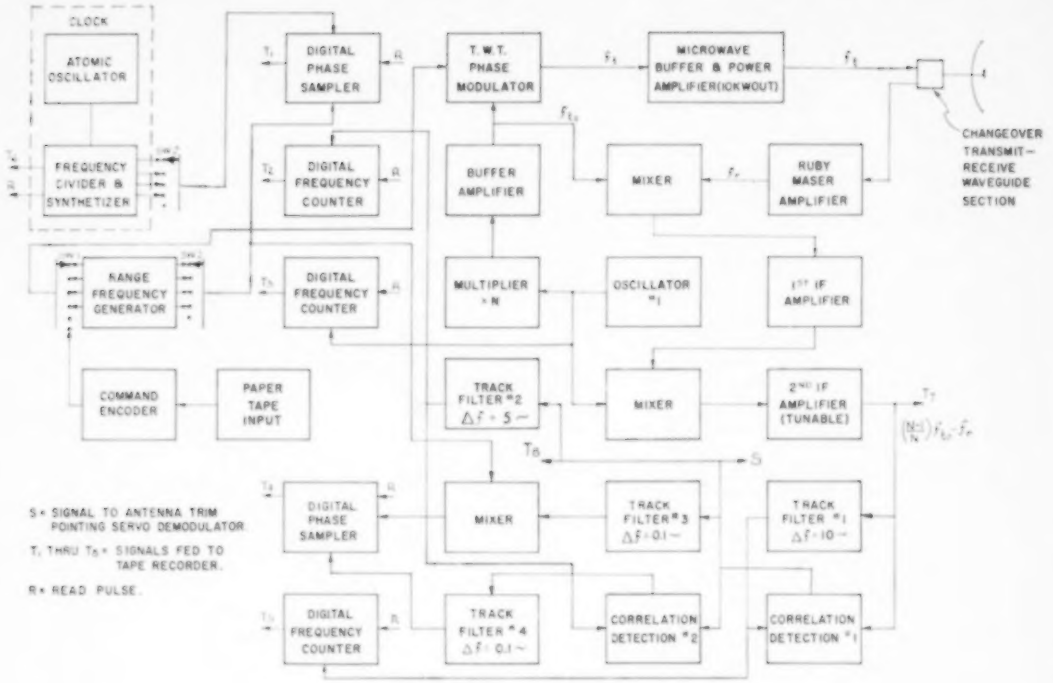


Fig. 5. Block diagram of the 2-10-8 ground station.

receiver i.f. amplifier is demodulated in the correlation detector (No. 1) at the ground station and filtered in the tracking filter (No. 2), then digitized and recorded on recorder channel T_2, thus completing all the measurements needed for Doppler ranging.

The output recorded on channel T_2 is also used to establish which frequency should be transmitted from the ground station in order to make the offset frequency in the transponder second i.f. amplifier fall in the center of the passband. Since this passband is only 100 c/s wide, the ground transmitter frequency must be adjusted very accurately to correct for Doppler shift and oscillator drifts.

At ranges up to a few hundred thousand miles, it is probably best to use automatic transmitter frequency adjustments. At these ranges response delays would be a maximum of a few seconds; however, as ranges in the millions of miles are reached it is probable that best results can be obtained by a skilled operator manually adjusting transmitting frequency with the aid of a precomputed schedule.

LIMITATIONS OF DOPPLER-ONLY MEASUREMENT

Doppler range measurements alone are ambiguous. In tracking spacecraft to ranges beyond a few thousand miles, intermittent tracking operation from any single ground station must be accepted. Even though there may be enough power available in the spacecraft so that the transponder can be allowed to transmit continuously, the Earth's rotation will prevent continuous communication.

Range determined by Doppler-only measurements from any one ground station would require that extrapolations of data be made over intervals without signal. Such extrapolations could be expected to introduce cumulating errors that seriously limit the accuracy of knowledge of the transponder range after many days of operation. Although additional ground stations

would help reduce this error, the likelihood of breaks in data is still high with an attendant loss in accuracy.

RANGING METHODS

In order to overcome the requirement for continuous data, it is proposed to use modulation frequencies superimposed on the carrier to allow range measurement with no unresolvable ambiguities. This brings up a technical problem of providing suitable receiving sensitivity while accepting signal sidebands which are spread over a relatively wide bandwidth.

Ranging sensitivity increases when higher ranging modulation frequencies are used. If the highest range frequency is doubled and the modulating index is held constant, the sensitivity of range measurement for a given signal-to-noise ratio in the data bandwidth is doubled. However, in a conventional receiver, doubling the modulation frequency at fixed modulation index would require twice the receiver bandwidth. For good accuracy, ranging modulation frequencies as high as 1 Mc/s are desirable.

Correlation detection techniques have been developed which allow operation of the frequency demodulator with input signal-to-noise ratios far below unity without further impairment of the signal-to-noise ratio. Post demodulator filtering can then be used to recover a reasonable signal-to-noise ratio in a narrow data bandwidth. Actual required data bandwidth for range measurement is less than 1 c/s. Such techniques overcome some of the limitations to increase of ranging modulation frequencies.

Correlation detection circuits are in general rather complicated and somewhat tricky in operation when applied to the problem of detection of signals far below the noise at their inputs. In general, they are more suited to use in ground equipment where careful adjustments can be made when necessary, than to use in a transponder that may have to operate a year or more unattended.

SIDEBAND FOLDING TECHNIQUE

A technique for folding sidebands, to allow bandwidth reduction before demodulation in ranging systems, has been under investigation at Convair (Astronautics) Division of General Dynamics the past few years. It appears this technique is particularly suitable for use in spacecraft equipment because of its inherent simplicity. The technique consists of modulating the local oscillator in a receiver with a slightly different modulating frequency than the modulation being received. This produces a clustering of sidebands as shown for one case in Fig. 6. The block diagram of a laboratory test setup and for demonstrating this principle is shown in Fig. 7. A photograph of oscilloscope pictures

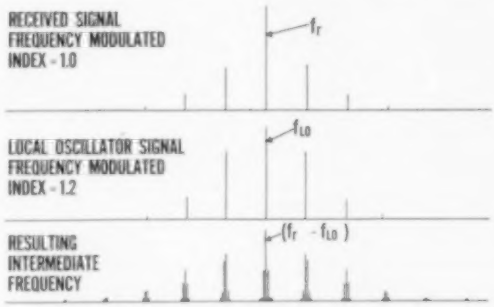


Fig. 6. Distribution of sidebands using sideband folding technique.

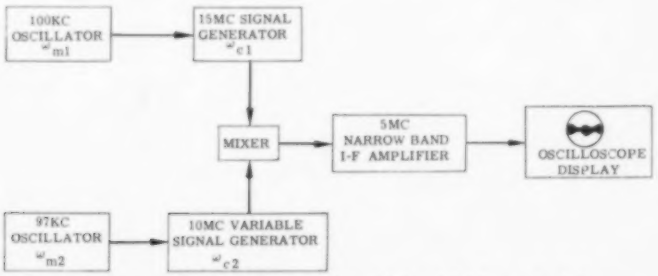


Fig. 7. Test setup for demonstrating sideband folding principle.

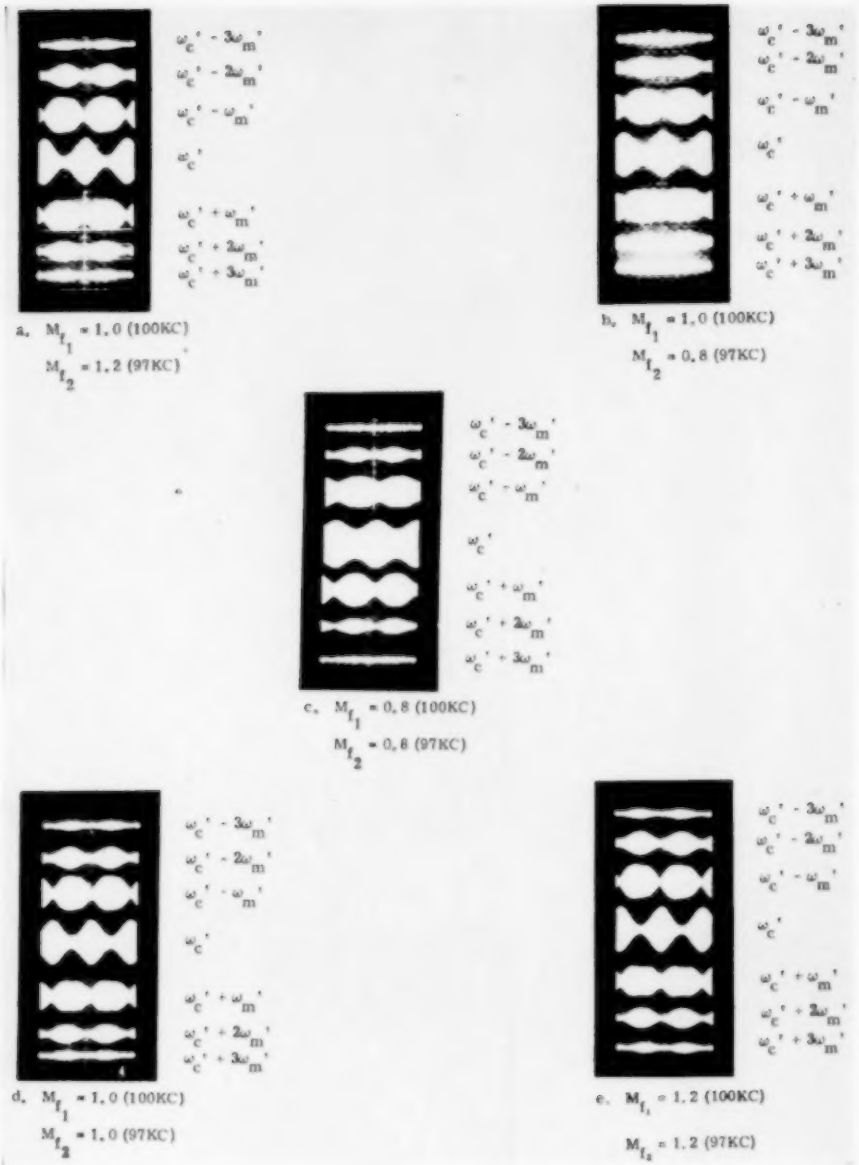


Fig. 8. Oscilloscope observations of signals formed by folded sideband clusters.

taken for several cases demonstrated with this equipment is shown in Fig. 8. Results are summarized in Fig. 9. The laboratory investigation was made by Mr. J. A. Webb.

It is interesting to note that if all the sidebands are filtered out, except the signals remaining clustered around the carrier of the mixed signal, the remaining signal resulting from combining two frequency modulated signals of low index and nearly equal modulation

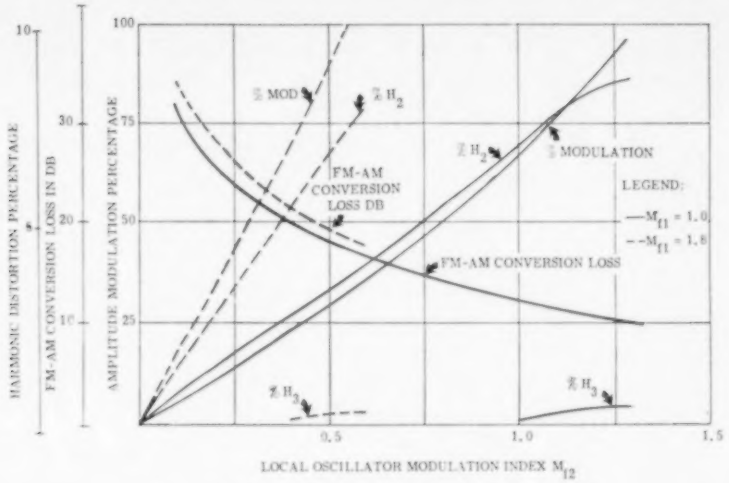


Fig. 9. Summary of results of laboratory investigation.

frequency is an amplitude modulation signal containing the frequency difference of the two modulating signals.

NON-AMBIGUOUS RANGING IN THE 2-10-8 SYSTEM

The sideband folding technique is used to add modulation ranging capability to the 2-10-8 System. The same crystal oscillator that feeds signals into the multiplying chain to provide frequency reference for the carrier signal transmitted by the transponder also is fed into a dividing chain (see Fig. 4). The input to the dividing chain and various frequencies tapped off this chain are fed to a sequencing switch which in turn is controlled by a programmer that receives its time reference from the dividing chain and command data. Each position (except one which is blank) of the sequencing switch allows modulation of the transponder with one of the frequencies from the dividing chain. By using a sequence of modulation frequencies, range ambiguity is resolved.

The signal in the second i.f. of the transponder is amplitude modulated by the frequency difference between the modulation frequency on the signal being received by the transponder and the modulation being transmitted. The ground station is adjusted to control the modulating frequency it transmits so that the modulation frequency will be about two cycles higher, when the signal reaches the transponder, than the modulation frequency being transmitted by the transponder. A typical maximum modulation frequency is about 1 Mc/s, thus Doppler frequency shift must be accounted for in establishing the ground station modulation frequency.

The amplitude modulation on the transponder carrier will have a frequency of about 2 c/s and will readily pass through the 100 c/s second i.f. amplifier bandwidth with very little phase distortion, even though it is not accurately centered in the passband.

At the ground station the signal received from the transponder is fed to suitable filters and correlation detectors.

A tracking filter (No. 3 in Fig. 5) with a bandwidth of $\frac{1}{10}$ c/s is used to filter out the signal produced by the ranging modulation generated in the transponder. Another tracking filter (No. 2) with a bandwidth of 5 c/s is used to filter out the carrier of the modulation signal imposed on the transponder.

The output of the tracking filter (No. 2) is used in determination of Doppler velocity as well as being fed to a correlation detector (No. 2) used to demodulate the approximately 2 c/s amplitude modulation on that signal. A tracking filter (No. 4) with a bandwidth of

$\frac{1}{10}$ c/s is used to track the 2-c/s output from the second correlation detector.

The output of tracking filter (No. 3) is mixed with the ground station ranging modulation frequency digitized and recorded on channel T_4. Since round-trip propagation time will become many minutes when tracking at 100 million mile ranges, the ranging frequency used in the comparison must be the ground station reference ranging frequency that was transmitted minutes earlier at the time the signal that is being received at the ground station was determined. There is no need to transmit continuously with a given modulation frequency until the return signal is received and measured, if the ground station operation is properly synchronized with the transponder. A single ground station can transmit then switch to receive, when range is great.

TRANSPONDER PROGRAMMER

The transponder is equipped with a programmer which will turn the transponder on at a preset time and sequence the transponder in accordance with commands from the ground station. In the event that no command is received the programmer will turn the transponder on and sequence it through the operating cycle at precisely determined times. Time for the programmer will be determined from the oscillator and dividing chain used for ranging modulation frequencies. The programmer, oscillator and dividing chain will be operated continuously. All other parts of the transponder will operate only on program command.

Cycling the transponder off and on with the programmer allows a great reduction in the total power requirements. This makes it possible to use a storage battery recharged by a relatively small number of solar cells for electrical power. In order to operate the transponder directly from solar cells, very many more over a much larger area would be required with the complication that the solar cells would become a serious cost, size and weight problem. Cycling also reduces the problem of keeping the transponder at a reasonable operating temperature. Cycling will make jamming or unauthorized triggering of the transponder more difficult and, if properly used, can increase reliability by reducing the total operating time. Many months will go by during the trip of a spacecraft to another planet when relatively few trajectory measurements are needed. Early in the flight, as well as when the target plant is approached, more frequent samples of data are required.

OPERATING MODES

The design of the 2-10-8 System is such that, even if the ground-to-transponder radio link should fail, the course of the spacecraft could be followed to a lesser accuracy. In this case, the range measurement would depend directly on the accuracy of the transponder and the ground station clocks.

When the transponder is operating against a single ground station with a single antenna, there are some operating limitations. In order to take advantage of the useful sensitivity of a maser amplifier between the antenna and the ground station receiver, there must be very low losses between the antenna and the maser. Any losses in this part of the system greatly increase the effective noise temperature of the receiving system. In the sensitivity region being considered, a loss of 2 dB between the antenna and the maser amplifier could increase the effective input noise temperature by as much as 8 dB. This would result in a total loss in receiving sensitivity of 10 dB, reducing receiving range to about one third.

Duplexing both transmitter and receiver simultaneously into one antenna would almost certainly entail appreciable loss and would also allow some noise from the transmitter to come into the receiving bandwidth, further limiting the receiving sensitivity. Thus, for maximum sensitivity, an antenna can only be used either for transmission or for reception at any one time. Since a large antenna suitable for use in this system would represent a major part of the station cost, techniques have been developed so that only a single antenna is required at each ground station. If desired a second antenna suitably separated from the first can be added at a later time to provide the

system with the greater flexibility of being able to transmit and receive simultaneously.

A remotely controlled mechanical waveguide switch is planned for switching the antenna with minimum transmission line loss. For tracking at ranges of hundreds of millions of miles, the transmission time is long enough to transmit for many minutes and then to switch and receive the signals returned by the transponder. At shorter ranges a small secondary antenna of low cost, mounted on the big dish, could be used for receiving.

OPERATING METHODS

In the actual case of tracking spacecraft, there will be large changes in range between the three pairs of measurements. The motion of the Earth is quite accurately predictable and the motion of the spacecraft is very smooth. Thus, by accounting for Earth rotation and motion around the Sun, the remaining motion is very slowly changing. By use of a series of repeated measurements, the motion can be precisely taken into account in reducing the observations. As suggested by Dr. R. M. Leger, motion of the Earth in effect provides a very long baseline for determining the direction of the transponder in the plane of the Earth's motion. Since Earth rotational and orbital motions are in different planes, a precise measurement of spacecraft range from a single station on the Earth could, in principle, provide a three-dimensional fix on the vehicle. This technique fails to yield accuracy in one dimension, however, when tracking in certain directions with respect to the directions of Earth rotation and orbit motion.

For a minimum 2-10-8 System installation, two ground stations, one located in high northern latitudes and the other in far south latitudes and both near the same longitude, are recommended.

There are both accuracy and reliability advantages to having one or more additional stations; however, it does increase the cost.

A logical program sequence might start with transmission from Ground Station 1 to the transponder with Ground Station 2 receiving. After sequencing through the modulating frequencies, Ground Station 1 would be converted to receive and record the transponder reference measurements. After Ground Station 2 had received a full sequence of ranging frequencies transponded while the transponder received Ground Station 1, Ground Station 2 would be converted to transmitting. Ground Station 2 would then program through the sequence of modulations, then switch to receive and the return signal would be received by both ground stations. There could be up to about half an hour delay depending on the distance of the spacecraft between the time the transponder received Ground Station 1 and Ground Station 2. The transponder could be programmed off for part of this interval and only be on during the time necessary to come up to operating temperature and return the full ranging sequence of transmission from each of the ground stations. Such a sequence would require about 10 min each time plus 1 min of warm-up.

SIGNAL POWER

Ground Station to Transponder Link

Ground station transmitter power (10 kW)	+40 dBW
Loss in isolator and waveguide to antenna	−2 dB
Ground station antenna gain (85 ft dia.)	+60 dB
Space attenuation (2×10^8 naut. miles at 5000 Mc/s)	−278 dB
Spacecraft antenna gain	+14 dB
Loss between antenna and transponder mixer	−1 dB
Loss due to sideband folding into narrowband (approx.)	−3 dB
Equivalent signal power in transponder i.f. amplifier bandwidth	−170 dBW
Effective receiver noise power (6 dB n.f. and 100 c/s bandwidth)	−178 dBW
Effective signal-to-noise ratio in 100 c/s bandwidth	+8 dB
Signal-to-noise ratio required for acceptable performance	+3 dB
Margin for acceptable performance	+5 dB

Transponder to Ground Station Link

Transponder power output (50 W)	+17	dBW
Loss between transponder and antenna	−1	dB
Effective gain of spacecraft antenna	+14	dB
Space attentuation (2×10^8 naut. miles at 5000 Mc/s)	−278	dB
Ground station antenna gain (85-ft dish, 5000 Mc/s)	+60	dB
Loss between antenna and maser amplifier	−0·1	dB
Signal power received at ground station	−188·1	dBW

Transponder Carrier Measurement

Transponder power available at receiver input in 10 c/s bandwidth of tracking filter (No. 1) in ground station (50 per cent transponder power assumed in carrier)	−191·1 dBW
Power required for satisfactory measurement (receiver equivalent noise temperature assumed 30 °K and required s/n assumed 6 dB)	−198 dBW
Margin for acceptable performance	+6·9 dB

Transponder Generated Subcarrier Phase Measurement

Signal power in these sidebands at ground station receiver input (20 per cent transponder power assumed in these sidebands)	−195·1 dBW
Power required for satisfactory measurement (bandwidth = 0·1 c/s, convolution loss = 3 dB and required s/n assumed 20 dB)	−201 dBW
Margin for acceptable performance	+5·9 dB

Transponder Carrier Offset Measurement

Signal power in these sidebands at ground station receiver input (30 per cent transponder power assumed in these sidebands)	−193 dBW
Power required for satisfactory measurement (bandwidth = 5 c/s, convolution loss assumed = 3 dB and required s/n assumed 6 dB)	−198 dBW
Margin for acceptable performance	+5 dB

Transponder Modulation Offset Measurement

Signal power in these sidebands at ground station receiver input (50 per cent of the power in the 5 c/s bandwidth containing carrier offset information data, i.e. 15 per cent of total transponder power)	−196 dBW
Power required for satisfactory measurement (bandwidth = 0·1 c/s, two convolution losses of 3 dB each and required s/n assumed 6 dB)	−196 dBW
Margin for acceptable performance	0 dB

SIGNAL MARGIN COMMENTS

The most critical measurement from a signal-to-noise standpoint is the measurement of the 2 c/s modulation offset transmitted by the transponder. It is critical because it requires detection two times resulting in 3 dB convolution losses each time and also because the energy in the 2 c/s sidebands is a very small part of the transponder transmitted power.

If ranges greater than 2×10^8 nautical miles are needed, these can be achieved by increasing antenna gain. Beyond 2×10^9 nautical miles, it would be wise to reduce the bandwidth of the tracking filter handling this data below the 0·1-c/s indicated here.

It should be noted that as long as tracking is accomplished without gain or loss of cycles and with a number of digital samplings per cycle, the final filtering to a much lower bandwidth can be accomplished on the output data. Reduction of tracking bandwidths is usually accompanied with increased operational complications; thus unnecessarily narrow bandwidth tracking should be avoided whenever possible.

A signal-to-noise ratio of 6 dB is considered necessary to avoid loss or gain of cycles in digitizing outputs. A higher signal-to-noise ratio is indicated as the acceptable data signal requirement for the transponder clock sideband measurements since it is easily attainable and since it is desirable to reduce data handling.

TRANSMITTING ON ONE GROUND STATION RECEIVING THE OTHER

The design of the ground stations is such that the transmitted carrier frequency and modulation phase signals are referenced against a precision atomic frequency-controlled reference clock when recordings are made. Similarly, the received carrier frequency and modulation phase signals are referenced to the clock during recording. This allows measurements to be made over a period of minutes and hours with very little relative drift between clocks.

This makes it possible to reference the signals transmitted from Ground Station 1 to Clock 1 and to record signals at Ground Station 2 referenced to Clock 2, then to reverse the process by transmitting from Ground Station 2 the signals referenced to Clock 2 and to record at Ground Station 1 the signals referenced to Clock 1. By repeating this process and accounting for known Earth motion, the relative time between Clock 1 and 2 can be determined as well as the transmittal time to the spacecraft and return. Since the spacecraft is also in motion, multiple observations are necessary to determine this

motion and to compensate for it. When recordings are made at the ground stations, the transponder signal contains information necessary to measure the clock setting in the transponder propagation time delay. By recording, at each ground station, this measurement simultaneously with the signal responded from the other ground station, the relative clock setting between all three clocks, the one at each ground station and the one in the transponder, can be determined.

RANGING

Once the relative time of the three clocks have been established, the sum and difference ranges and the range from either ground station can be computed. The sensitivity of measurement of nonambiguous range is of the order of 10 ft; however, here the error in the knowledge of the velocity of propagation, c, becomes a major limiting factor in converting the transit time into feet of range. Assuming the value of c is known to one part in 10^6, at 200 million nautical miles an error of 200 miles in range would result. This is a fundamental error common to all radio tracking systems. This error does not affect the ability of the system to relate distance to transit time which is more fundamental. During observations a measuring sensitivity to changes of spacecraft range of 0·1 ft is possible in principle; however, a frequency drift of about one part in 10^{13} can cause a similar indication and is not necessarily detectable. The carrier measurement is very useful in detecting target planet perturbations as early as possible. Since there is no prior experience in receiving precise ranging signals over similar ranges, it is not unlikely that some new factor influencing propagation may be discovered.

DIRECTION COSINE MEASUREMENT

If two stations separated by a straight line through the Earth of 20,000,000 ft in length (about 3,000 nautical miles) are used, an error of 10 ft in determining the difference of range would result in an error of 10 divided by 20,000,000 or 5×10^{-7} of the cosine of the angle between the direction of the spacecraft from a point midway between the ground stations and the line between the ground station.

An inaccuracy of this amount would contribute an error, in determining a spacecraft's location at a range of 200 million nautical miles, of 100 nautical miles in a direction parallel to the line between the antennas.

FREQUENCY

Any frequency over a range of a few hundred megacycles to many kilomegacycles could be used for transmission using the techniques outlined here. Since both carrier Doppler and ranging modulation signals are used, it would be possible to compensate for large errors caused by ionization. The carrier signal propagates at the phase velocity rate and the modulation signal at the group velocity rate; thus, the mean of these velocities is equal to propagation velocity without ionization. However, since continuous tracking is not possible, it is difficult to make this correction very precise.

In the 2-10-8 System, the frequency of 5000 Mc/s is chosen for illustration. This frequency is high enough to avoid significant errors from the anticipated ionization levels and high enough to avoid susceptibility to solar flares without being so high as to be impracticable. It is also a frequency at which galactic noise is at a minimum and ruby masers work well. At this frequency, operation to within a degree or so of the direction of the Sun appears feasible. It may be important to operate within a reasonably small angle of the Sun on some space-probe trajectories to Venus.

COMMAND

The link from the ground to the missile is suitable for carrying commands. Subcarriers of less than 50 c/s and over 4 c/s can be handled by the link. A system feature is that the command is automatically repeated as received by the transponder. One command channel is planned for resetting the transponder programmer when desired. Others could be made available.

TELEMETER

Telemetering subcarriers can be imposed on

the transponder signal in the spacecraft. Only one subcarrier at a time is planned and the bandwidth of the data on the subcarrier when operating at maximum tracking range should be under 10 c/s, although the subcarrier frequency can be up to several thousand cycles without causing serious receiving problems. It is planned to use the direct tape recording of the i.f. signal for post measurement evaluation. No real-time telemetering reduction in the ground station is planned; however, it could be added to the design.

SPACECRAFT ANTENNA POINTING

Beyond ranges of about 10^8 nautical miles, increasingly narrow beam antennas are required on the spacecraft. One plan would be to use a spinning vehicle with the antenna pointing very slightly off the spin axis so that the signal is conically scanned by the vehicle rotation. This allows a measure of the relative direction of the signal with respect to the vehicle spin axis. One mode of operation is to use the transponder-received signal to actuate tiny trimming jets to precess the spin axis until it is along the line to the ground station. A second method would be to operate the trimming jets by command, using the Sun as an additional direction reference by telemetering the signal from a solar cell mounted on the side of the vehicle and by using the solar cell for roll direction reference for applying the commands. For operation at 2×10^9 nautical miles, a 34 dB vehicle antenna is needed. Such an antenna would be about 5 ft in diameter and have a beamwidth of about 3°.

This technique can also be used for determining vehicle orientation for commanding a rocket thrust burst for orbit correction. The radio-determined direction and Sun direction can provide the necessary references. If the method of applying the correction requires changing the direction of the vehicle spin axis, a gyro would be needed to provide a temporary memory of the radio-determined direction to program the spin axis back after the manœuvre.

GUIDANCE PROBLEM

In guiding a spacecraft to a near planet, either to strike it, to land on it, to pass by very close or to orbit it, there are two major problems, even though the vehicle could be controlled to achieve an exactly predictable cutoff position and velocity.

First, the inaccuracy in knowledge of the Earth's mass can cause an error in spacecraft orbit as the spacecraft clears the near vicinity of the earth. The 2-10-8 System provides precise tracking data necessary to compute trimming thrust necessary for corrections. It also provides a command link over which commands can be sent to the vehicle to perform the needed corrections. The system can also be used to determine whether or not the corrections were satisfactory and, if necessary, to command additional corrections. These corrections should be applied within the first few weeks of flight to minimize rocket-thrust requirements.

The 2-10-8 System is suitable to use for interplanetary post-injection guidance. Mr. Maxwell Noton, JPL-NASA, presented a paper on the subject of post-injection guidance at a meeting in San Diego in 1959. In addition, the 2-10-8 System provides sufficient tracking range and accuracy to avoid the need for additional terminal guidance equipment for most applications.

The second major problem can cause a miss as great as 10–20 planet diameters for Venus and Mars probes. This problem is the inaccuracy in the knowledge of the astronomical constant (the Earth's distance from the Sun), which results in lack of exact knowledge of how far it is to the target planet. The 2-10-8 System provides sufficient tracking accuracy and sensitivity to measure perturbations in the course of a spacecraft as it approaches a target planet, while it is still far away.

From observations of the perturbations, the relative position of the spacecraft and the target planet can be established early enough to allow trimming velocity corrections to be applied to manœuvre the spacecraft to the desired course. Several corrections to the spacecraft as it approaches the target planet may be necessary to achieve accuracy with minimum rocket thrust. As in earlier orbit corrections, the commands to trimming rockets would be transmitted over the 2-10-8 System.

Control of the impact point or closeness of approach to an accuracy on the order of a few hundred miles appears reasonable.

Once the relative position of the target planet and the spacecraft have been well established (best if the spacecraft is put in orbit around the target planet), the measurement of range of the spacecraft will also provide a measurement of the range of the target planet. This will allow a very precise determination of the astronomical constant in terms of the velocity of radio propagation.

SATELLITE TIME REFERENCE

A satellite equipped with a transponder operating on the principle of the transponder of the 2-10-8 System could be used as an extremely accurate time reference. The orbit could be determined accurately by the 2-10-8 System or a lower power short-range version, in order to allow accurate corrections for transmission time and changes in transmission time.

A microwave receiver could receive both the transponder clock signals as well as the offset frequency relating it to higher accuracy ground station clocks. It is estimated that if a version of this system were utilized for this purpose, relative, time between ground stations any place on the Earth could be established to within a small fraction of a microsecond.

CONCLUSIONS

Techniques described in this paper will allow the design of a tracking, guidance and communications system with the following features:

(1) Guidance and execute commands can be transmitted to the space vehicle over the tracking link.

(2) Narrow band telemetering can be transmitted over the tracking link from the space vehicle to Earth.

(3) Two ground stations, with one large antenna each, can provide accurate three-dimensional tracking of a spacecraft transponder.

(4) Failure of the ground-to-vehicle radio link reduces tracking accuracy but some track data could still be obtained.

(5) The transponding of range in the vehicle does not require detection; thus complex correlation detection circuits are not used in the transponder. A simple narrow band detector is used for decoding command signals.

(6) Power is conserved in the vehicle by operating the transponder under programmer control only when needed.

(7) Guidance accuracy to a point on or near the target planets on the order of a few hundred miles is considered reasonable without use of additional terminal guidance.

(8) These techniques can be extended to the design of a time standard satellite which can establish relative time over the Earth to an anticipated accuracy of better than 10^{-6} sec.

(9) Operation to ranges of 2×10^9 nautical miles (twice the distance of Saturn) appears reasonable with reasonably high gain vehicle antennas. Vehicle stabilization spin axis along radio line of sight is assumed.

A VERTICAL AZIMUTH TRANSFER SYSTEM

CLAUDE L. EMMERICH and ROBERT E. WISNIEFF
Norden Laboratories, United Aircraft Corporation

Abstract—A polarimeter is described which is applied to the problem of transferring azimuth information in a vertical direction. High precision and automatic operation are achieved in the polarimeter by modulating the plane of polarization in a Faraday Cell. The principal advantage derived from this type of transfer system is that the line of sight remains fixed independent of the orientation of the reference direction being transferred.

Applications include platform-mounted polarizers, frame-mounted polarizers, and a manually adjustable alignment system. Performance data indicate a precision of 5″ of arc at a vertical separation of 40 ft.

1. INTRODUCTION

The stabilizing properties of gyroscopes used in guidance systems are characterized by the fact that the output signal of a gyro represents the deviation angle from a reference direction fixed in inertial space. The orientation of this reference direction in space is completely arbitrary as far as the gyro is concerned. In order to derive useful information from gyroscopes in solving navigation or attitude control problems, it is therefore necessary to provide initial orientation signals to the gyros. The process of imposing the initial orientation information on the guidance system may be performed in a variety

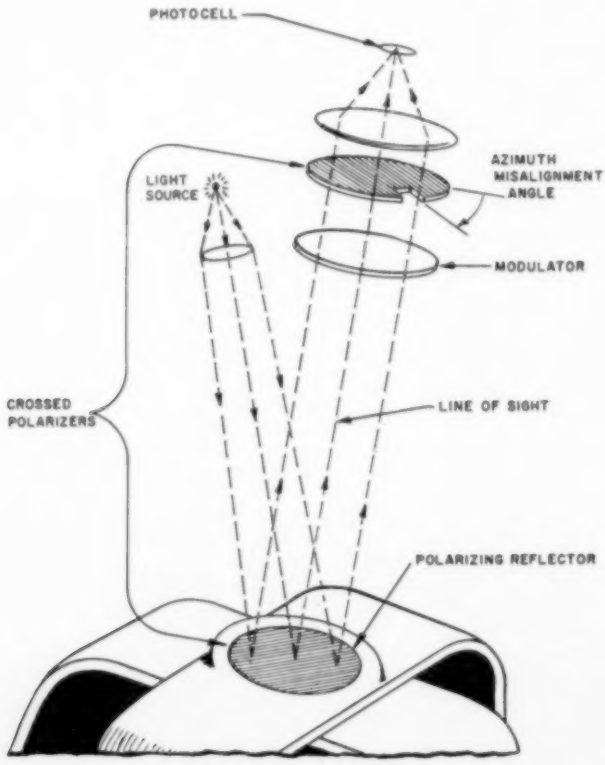


Fig. 1. Polarimeter system.

of ways, including caging to a mechanically oriented framework, erecting to the vertical as dictated by pendulum or acceleration signals, slaving in azimuth to magnetic signals or aligning by optical means to benchmarks or stars.

The vertical azimuth transfer system described in this paper belongs to the class of optical alignment systems. It is distinguished from other optical systems by the property of measuring one horizontal plane to another is accomplished by referring to the plane of polarization of a beam of light transmitted in a vertical direction. Thus the basic instrumentation required is essentially a polarimeter. The same technique can of course be utilized in transmitting rotational information along lines of sight that are not vertical; however, in most missile alignment applications roll and pitch

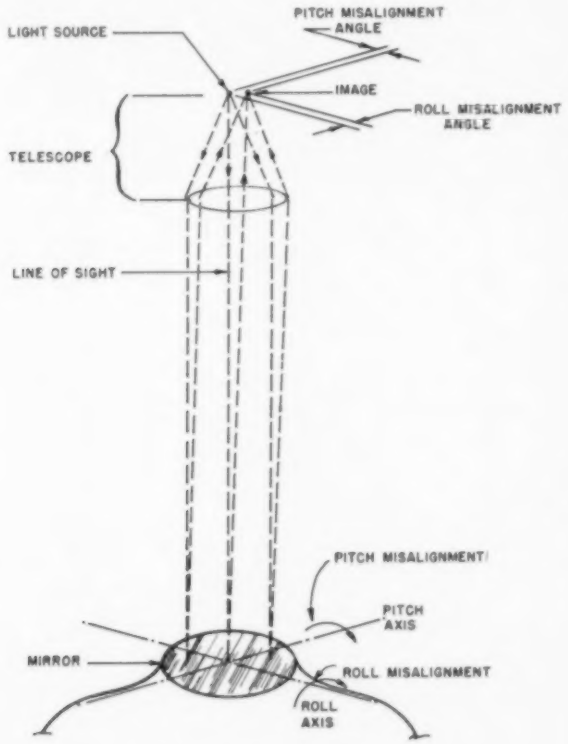


Fig. 2. Autocollimator system.

angular rotations about the line of sight in contrast with the familiar autocollimating system which measures rotations about axes at right angles to the line of sight. It thus complements the autocollimating system so that the alignment is completely definable by transmission of optical information along a single line of sight, or at least parallel lines of sight, as shown in Figs. 1 and 2.

The transfer of azimuth information from information is readily available internally from gravity-sensing accelerometers or pendulums whereas the azimuth information must be provided by external means.

The vertical azimuth transfer system using polarized light has three major advantages over some other optical or mechanical azimuth alignment systems.

In the first place the system can be easily calibrated and used over great variations in

range of vertical separation. Since the plane of polarization of the beam is not measurably rotated by passing through air even over very great distances, an alignment calibration carried out at essentially zero distance of separation is preserved everywhere along the entire useable length of the line of sight.

Secondly, the lateral alignment of the transmitting and receiving stations is permitted to vary over wide ranges. It is only necessary that a part of the aperture of the receiver is located inside the cone of light emanating from the transmitter.

Third, the alignment angle direction is not limited. That is, any desired orientation of the azimuth reference direction can be transferred without changing the line of sight. For example, the directions of true north, the direction to a bench mark, or different initial directions of great circle courses to a given set of terminal areas can all be transferred in a single setting of the instruments.

A minor advantage which might be of importance in some applications is that the system can be calibrated directly without any additional instrumentation by reference to the plane of polarization of light from the sky, which is oriented at right angles to the azimuth of the sun.

A disadvantage of the system is that any windows interposed in the path of the polarized light beam must be made of strain-free material to prevent rotation of the plane of polarization or at least to assure that the rotation is constant so that it can be compensated.

2. BASIC INSTRUMENTATION

In order to achieve a high degree of precision in this optical data transmission system the selection of the instrumentation is critically dependent upon a number of factors such as light intensity, drift, polarizer accuracy, wear of moving parts and temperature variation effects.

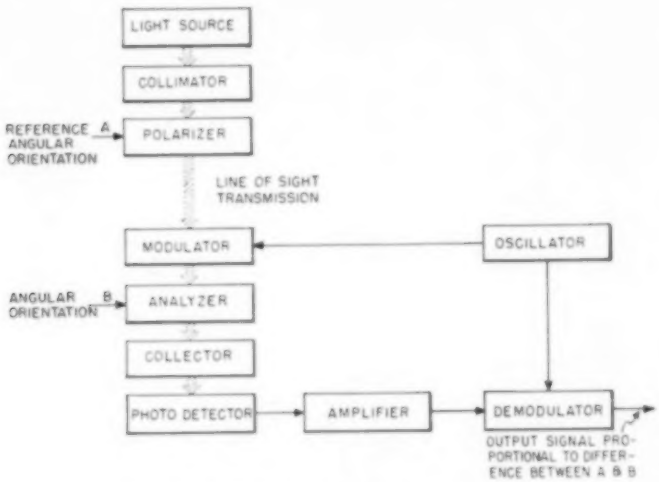


Fig. 3. Block diagram of polarimeter system.

The block diagram of a vertical azimuth transfer system designed to operate with high precision is shown in Fig. 3.

The light from an incandescent filament bulb is collected by a collimator and directed towards a polarizing prism which is rigidly attached to the mechanical structure aligned to the reference direction. The plane of polarization of the light emerging from the polarizer then carries the reference direction along the line of sight towards the receiving unit. After passing through a modulator, which will be discussed later, the polarized light is sent through a second prism, called the analyser, which is rigidly

attached to the mechanical structure aligned to the direction whose angular orientation is to be measured. If the polarizer and the analyser are "crossed", a minimum of light will pass through the analyser. This light is collected and focused on a photo detector where it is transformed into an electrical signal. Subsequent amplification and processing of the signal then serves to indicate whether the polarizers are crossed.

In order to provide a signal that can be used not only to determine whether the polarizers are crossed but also to sense the direction in which to turn either polarizer towards a better null, a modulating–demodulating scheme is used. The plane of polarization is oscillated about the line of sight between the two polarizers so that the photo detector signal alternates. The phase relationship between the modulator drive signal and the photo detector output contains the information necessary to derive a direction-sensitive output signal which can then be used for the alignment of the equipment to which the polarizers are attached. The operation of the system will be evident from the following analysis.

2.1. Analysis of the System

The illumination transmitted by a pair of polarizers is given by the relation

$$I = I_0 \sin^2 \theta + I_0 e \qquad (1)$$

where I = transmitted illumination
I_0 = incident illumination
θ = angle between plane of polarization and analyser
e = residual transmission factor.

The modulator introduced in the optical path will cause a harmonic oscillation of the angle of the plane of polarization, as indicated in the following equation:

$$\theta = \theta_e + \phi_0 \sin wt \qquad (2)$$

where θ_e = angular offset of polarizers
ϕ_0 = peak amplitude of modulation
w = angular frequency of modulation
t = time.

Substituting equation (2) in equation (1) and replacing the sine of the small modulation angle by the argument yields the expression

$$I = I_0(\theta_e + \phi_0 \sin wt)^2 + I_0 e \qquad (3)$$

The expansion of the squared term results in the equation

$$I = (I_0 e + I_0 \theta_e^2 + \tfrac{1}{2} I_0 \phi^2) - \tfrac{1}{2} I_0 \phi_0^2 \cos 2wt + 2I_0 \theta_e \phi_0 \sin wt \qquad (4)$$

The first group of terms in equation (4) is independent of the modulation frequency and may be regarded as a direct component, consisting of the residual illumination, the square of the polarizer misalignment angle and the amplitude of the modulation.

The second term is a second harmonic component which is removed by filtering or demodulating.

The last term represents the a.c. fundamental component which discriminates directly in a phase-sensitive manner between various misalignment angles θ_e.

Three curves of the significant terms of equation (4) are plotted in Fig. 4 for different values of θ_e. The parabolic characteristic corresponds to the polarizer transmission response near null. The harmonic modulation is plotted on the ordinate and the resulting light transmission on the abscissa for the cases θ_e negative, $\theta_e = 0$ and θ_e positive.

In case (a) where $\theta_e = 0$, only even-harmonic components are present. In cases (b) and (c) where the error angles are of opposite sign, the harmonic components are of equal amplitude but the fundamental components are reversed in phase. Demodulation with reference to the fundamental component therefore results in an output signal proportional to the coefficient $2I_0\phi_0\theta_e$.

It is seen that the amplitude of the output signal is affected by the intensity of the incident illumination and by the amplitude of the modulation.

The derivation above is only valid for perfect harmonic modulation. Imperfections in the modulator cause errors that appear as jitter or drift in the signal. For high precision polarimeters, it is therefore important to provide a closely controlled modulation signal.

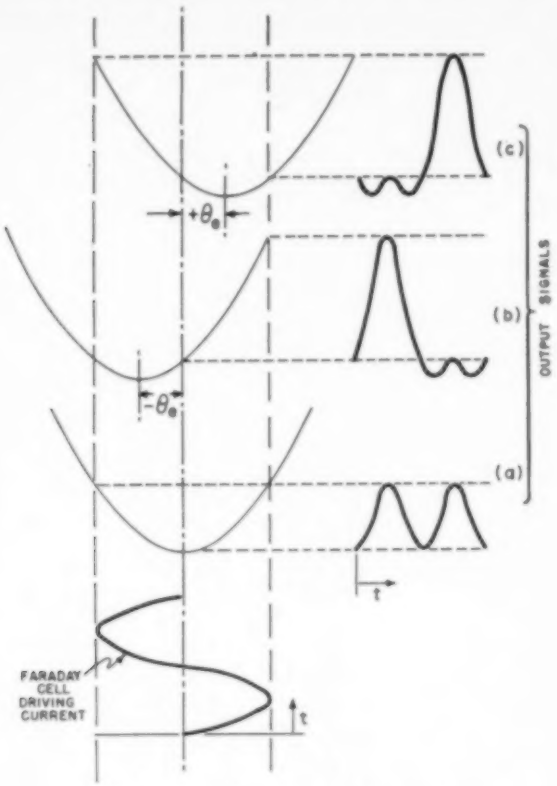


Fig. 4. Photo detector output signals.

2.2 Special System Components

2.2.1. *Light Source and Collimator*

The light source and collimating optics must be so designed that sufficient luminous intensity is received at the remotely located photo detector. Since the polarizer pair is operated near extinction, the intensity is reduced to a small fraction of the illumination which would be received in the absence of these elements. A concentrated high brightness source and optical system of minimum divergence are therefore required. For best operation, the light source should also be geometrically stable and have a minimum of noise component at the modulation frequency.

2.2.2. *Polarizers*

The polarizer pair must be carefully selected to give adequate system accuracy. Polaroid elements are desirable since they may be obtained in large cross section and hence permit an optical system with large aperture and high brightness at the detector. However, measurements made on several polaroid samples show repeatable variations in the plane of polarization in the order of 1′ of arc for various areas of the same sample. For systems requiring accuracy in the order of 1″ of arc this material would have to be carefully selected and checked at frequent intervals since exposure to a high intensity source causes a gradual deterioration of the polarizing properties.

Carefully cut natural crystal, such as Glan Thompson calcite prisms, show excellent uniformity in the plane of polarization but are somewhat limited in both aperture and field angle and consequently contribute to the design problems of the optical system. In some applications the better accuracy of the prisms may have to be sacrificed in favor of the greater aperture, increased field angle and flatness of polaroid sheets.

2.2.3. *Modulator*

The modulator in this system utilizes the Faraday effect, which is the rotation of the plane of polarization suffered by a beam of light passing through certain transparent media in a direction parallel to a magnetic field. The rotation is proportional to the magnetic field intensity and the path length, as expressed in the equation

$$\phi = VHd \tag{5}$$

where ϕ = rotation of plane of polarization
V = Verdet constant
H = magnetic field intensity
d = path length common to magnetic field and transparent medium.

If alternating current is applied to the solenoid generating the magnetic field, the magnetic field intensity varies harmonically:

$$H = H_0 \sin wt \tag{6}$$

where H_0 = amplitude of magnetic field intensity.

Substituting equation (6) in equation (5) results in the desired modulation of equation (2):

$$\phi = VH_0 d \sin wt \tag{7}$$

The identity with equation (2) is completed if the modulation amplitude ϕ_0 is defined by the relation

$$\phi_0 = VH_0 d \tag{8}$$

A suitable modulator thus consists of a coil of wire through which alternating current flows, and which has a core of dense flint glass with a high Verdet constant. This assembly is the familiar Faraday Cell. The line of sight passes along the axis of the solenoid assembly through the glass. According to equation (7) the rotation is in phase with the magnetic field which is in phase with the current through the coil. Hence the modulation reference signal required for the phase sensitive demodulation of the photodetector signal is derivable from the current through the Faraday Cell.

The Faraday Cell as a modulator is preferred for this application because it has no moving parts that could introduce jitter or suffer from problems associated with mechanical motion such as wear or lubrication difficulties in operating at the extreme limits of a given temperature range. Variations in temperature have to be considered, however, because the strain pattern in the glass of the cell affects the operation of the device.

Since the sensitivity of the system increases linearly with the polarization modulating amplitude, every effort must be made in the design of the Faraday Cell to produce maximum field intensity in the magnetic solenoid. A solenoid wound with copper tape and mylar insulation has been chosen to produce a strong field with minimum power input. Since the output information rate is determined by the frequency applied to this high-Q coil, rather high voltages may exist across the coil at the frequencies employed. Hence the insulation characteristics must be designed with care.

The glass used to produce the Faraday rotation must not only have a high Verdet constant but must also be free of strains to minimize depolarizing components. Effects of internal reflections within the glass should also be minimized. Coating the sidewalls with a material such as "Luxorb" paint will eliminate these reflections.

2.2.4. *Photo Detector*

The photo detector must be carefully selected to give maximum signal-to-noise ratio under low light levels. A photomultiplier tube was found to be the best detector. However, when working with a tungsten filament source, the sensitivity of a lead sulfide detector is as great as that of the photomultiplier. The lead sulfide detector may be preferred for a number of applications since it is smaller and requires less complicated associated circuitry.

The lead sulfide detector also attenuates the leakage at the blue end of the visible spectrum, which is an advantage when polaroid elements are used. To obtain maximum signal-to-noise ratio it is desirable to choose a detector of minimum size and to use a collecting lens to concentrate the illumination emerging from the analyser.

2.2.5. *Electronics*

In the electronic circuitry it is obviously essential that the amplifier have stable gain and phase shift characteristics. The phase sensitive demodulator is required to have an extremely high even-harmonic rejection ratio and have sufficient dynamic range that the even-harmonic components at null will not overload it. It should be observed that the even-harmonic amplitude increases as the square of the modulation amplitude so that while increasing the modulating amplitude increases system sensitivity the demands on the phase sensitive demodulator are also increased. Since drift in the output of the phase sensitive demodulator is intimately associated with the dynamic range, care must be exercised in selecting the modulation amplitude to assure that drift will be negligible with respect to minimum output. In experiments performed to date the modulation amplitude has not been large enough to cause any problems due to this effect.

3. APPLICATIONS

The foregoing discussion has established the basic principles of operation and some of the special component design requirements which must be observed to achieve a high precision system. The number of applications is of course unlimited and it is only possible to present a few typical configurations that may be useful in azimuth alignment or azimuth angle measuring techniques.

In its most precise applications, the vertical azimuth transfer system is used as a nulling device, merely indicating whether the orientation of one of the polarizers is at right angles to that of the other within a few seconds of arc or minutes of arc depending upon the precision and noise limitation characteristics built into the equipment.

It is, however, also possible to perform angular deviation measurements because the current which modulates the orientation of the polarization plane in the Faraday Cell can also be used to insert a direct-component offset angle of known magnitude. This application is sketched in the block diagram of Fig. 5 showing a high-grain feedback loop which automatically adjusts the direct component of the Faraday Cell current until a null is obtained. The d.c. component of the Faraday Cell current then is a

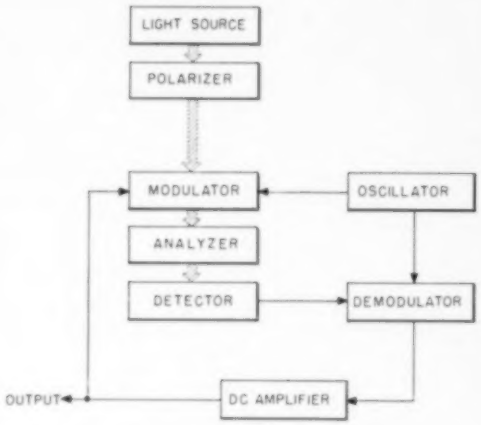


Fig. 5. Block diagram of angle measuring system.

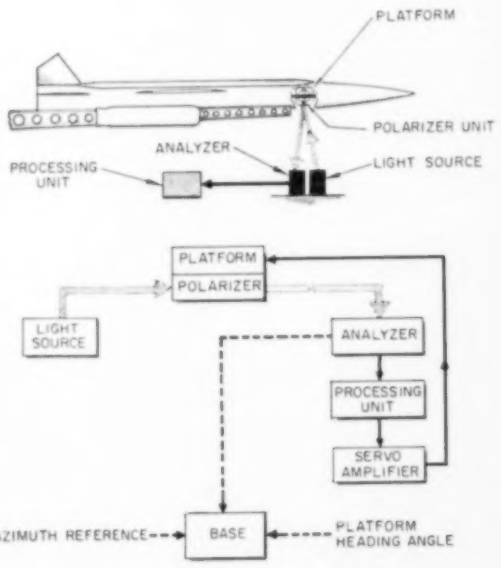


Fig. 6. Platform-mounted system.

linear measure of the angle of deviation between the two polarizers.

Three applications of a nulling vertical azimuth transfer system are shown in Figs. 6, 7, 8 and 9. From the point of view of precision, the system

in Fig. 6 is preferred. One of the polarizers is mounted directly on the stabilized gimbal of the gyro platform to be aligned. Its orientation is compared with the orientation of the analyser which reorients the analyser until the alignment error vanishes. When this condition is attained, the heading of the missile frame has in effect been transferred to the orientation of the analyser

Fig. 7. Frame mounted system, pictorial.

on the ground. The resulting electrical signal from the processing unit is amplified and used to precess the platform (with its attached polarizer) until the alignment error vanishes.

When this condition is attained, the orientation of the platform in azimuth matches that of the analyser, which has previously been aligned to the desired platform heading. The desired platform heading is usually defined by the platform heading angle, measured from an azimuth reference such as geographic north, which is introduced into the base of the system by star sights, benchmarks or magnetic references.

Somewhat less expensive but also less accurate is the system illustrated in Figs. 7 and 8, where one of the polarizers is attached to the outside of the missile frame. In this system the orientation of the missile frame is compared with that of the analyser on the ground. The electrical signal resulting from this comparison operates a servo

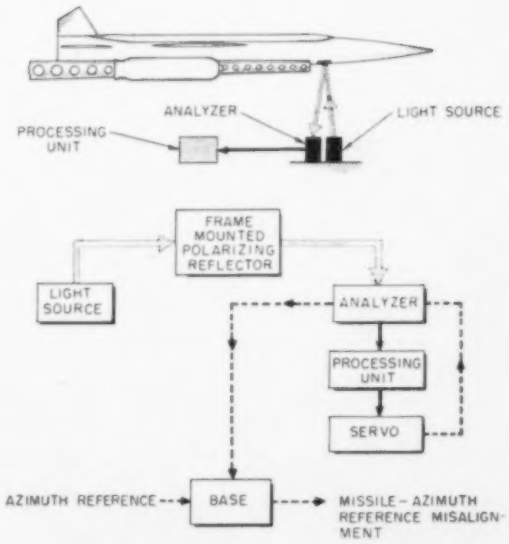


Fig. 8. Frame mounted system, schematic.

on the ground, where the heading angle can then be measured with reference to a previously established azimuth reference such as geographic north. The output from this alignment system is the angle between the missile frame and the azimuth reference. This signal would normally be used to precess the guidance platform inside the missile until the heading angle read by the platform is equal to the heading angle of the missile frame. The frame-mounted system is inferior to the platform mounted version of Fig. 6 because the movements of the missile frame on the launcher, which may be rather severe in the case of flexible structures in a high wind, are processed in the loop. These movements are bypassed in the platform mounted system by the platform's stabilizing characteristics.

In a third application (Fig. 9) the azimuth information is transferred vertically in a gantry structure to the level of the guidance platform in the nose of the missile. A standard autocollimator then provides the transfer link in the horizontal plane.

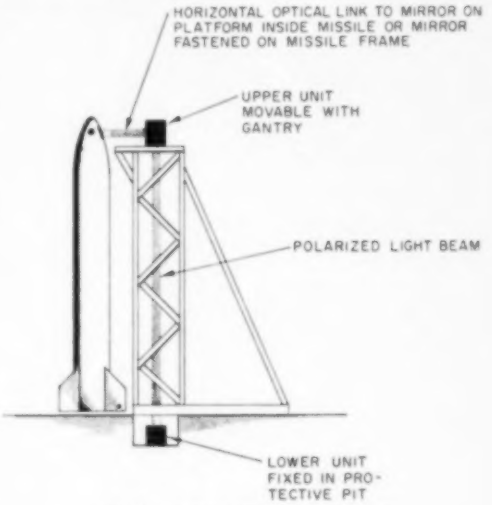

HORIZONTAL OPTICAL LINK TO MIRROR ON PLATFORM INSIDE MISSILE OR MIRROR FASTENED ON MISSILE FRAME
UPPER UNIT MOVABLE WITH GANTRY
POLARIZED LIGHT BEAM
LOWER UNIT FIXED IN PROTECTIVE PIT

Fig. 9. Vertical launch application.

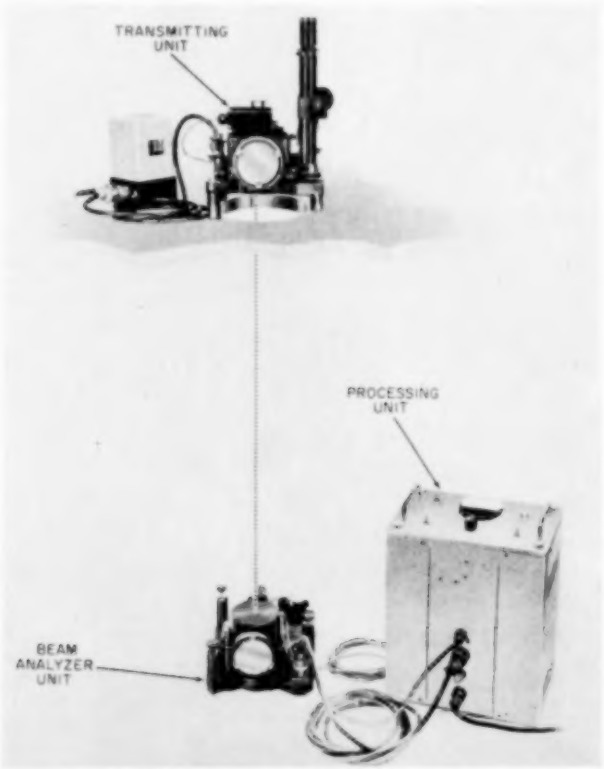


Fig. 10. New York Naval Shipyard alignment system.

Two systems employing the polarizing technique described here have been constructed to date at Norden Laboratories. One of them was designed as a high precision optical angle pickoff in a gravity meter constructed under contract NOrd 17421 for the Bureau of Ordnance of the Department of the Navy.

The other system has been constructed under contract N140(131)61894B for the New York Naval Shipyard, for use in transferring an azimuth reference direction over a vertical distance of 40 ft from one deck to another on the USS Compass Island. Fig. 10 is a photograph of the equipment delivered to the Navy Yard. Mirrors are used for transferring the azimuth reference directions in the horizontal planes at the upper and lower stations, manual adjustments being provided to achieve the required null. The electronics processing unit consists of a preamplifier, demodulator and d.c. amplifier. The output is a meter indication of the error signal. Three ranges of sensitivity can be selected by a range switch, nominally adjusting the full scale deflection of the meter for the ranges $-50-+50''$ of arc, $-10-+10'$ of arc and $-1{\cdot}5-+1{\cdot}5°$.

Performance tests of this instrument prove that over a 40-ft optical path length an azimuth alignment accuracy within 5″ of arc can be achieved.

GENERALIZED TRAJECTORY CHARTS

A. D. COHEN
Rocket Engine Section, Flight Propulsion Laboratory Department
General Electric Company, Malta Test Station

Abstract—A rapid method of obtaining trajectory information for preliminary design purposes is presented in graphic form. The effects of thrust variation with altitude and drag can be approximated by a simple iterative solution. By use of the curves data may be obtained for velocity, path angle, altitude and range covered during powered flight. Curves have been obtained for gravity turn flights and constant attitude flight paths, wherein the thrust vector angle remains constant relative to the local horizontal. The use of these curves is illustrated for a ballistic missile range calculation.

Digital computers and desk calculations represent the two extremes utilized in obtaining trajectory data. Digital computers will provide highly accurate data, but they are not always readily accessible or, in the case of preliminary design work, the required accuracy might not justify the cost of obtaining the digital data. Desk calculations, of course, are laborious and time-consuming for the step-by-step solution. The standard burnout velocity equation will yield fairly accurate velocity information, but gives no information relative to path angles and is somewhat inaccurate in estimating the effects of drag and thrust variation with altitude. The method described in this study represents the middle ground, whereby accurate data can be generated quite rapidly with the trajectory charts.

A family of charts was obtained from an IBM 704 digital computer trajectory program; several typical charts are illustrated in Figs. 1–9. The velocity and altitude data are presented in dimensionless form to facilitate their usage. Fig. 1 presents data to account for the variation of thrust with altitude as a function of chamber pressure, nozzle expansion ratio and average altitude. Figs. 2–4 incorporate the drag loss and Figs. 5–7 present data on velocity, altitude and range for a gravity turn trajectory. The range data is useful in ballistic missile analysis to evaluate the range covered during powered flight. Fig. 8 gives velocity data for the constant attitude type of trajectory. Note that an initial velocity is specified for this curve. In Fig. 9 the free flight range is given as a function of burnout velocity and burnout angle, assuming no re-entry drag.

SPECIFIC IMPULSE RATIO (FIG. 1)

This curve was obtained by application of the data presented by Bowlin and Dresser.[1] From this reference, the value of thrust coefficient, C_F, at sea level can be expressed as

$$C_{F_{S.L.}} = C_{F_{Vac}} - \frac{e_z}{P_C} P_A f \qquad (1)$$

At sea level, $P_A = 14{\cdot}7$, and equation (1) may be written as

$$C_{F_{S.L.}} = C_{F_{Vac}} - \frac{15{\cdot}72 e_z}{P_C} \qquad (2)$$

Using equation (1), values of $C_{F_{Vac}}$ vs. e_z were plotted on log–log paper and the following expression resulted

$$C_{F_{Vac}} = 1{\cdot}485 e_z^{\,0{\cdot}0614} \qquad (3)$$

Thus equation (2) may now be expressed as

$$C_{F_{S.L.}} = 1{\cdot}485 e_z^{\,0{\cdot}0614} - \frac{15{\cdot}72 e_z}{P_C} \qquad (4)$$

Also,

$$C_{F_{Avg}} = C_{F_{Vac}} - \frac{1{\cdot}07 e_z P_{Avg}}{P_C} \qquad (5)$$

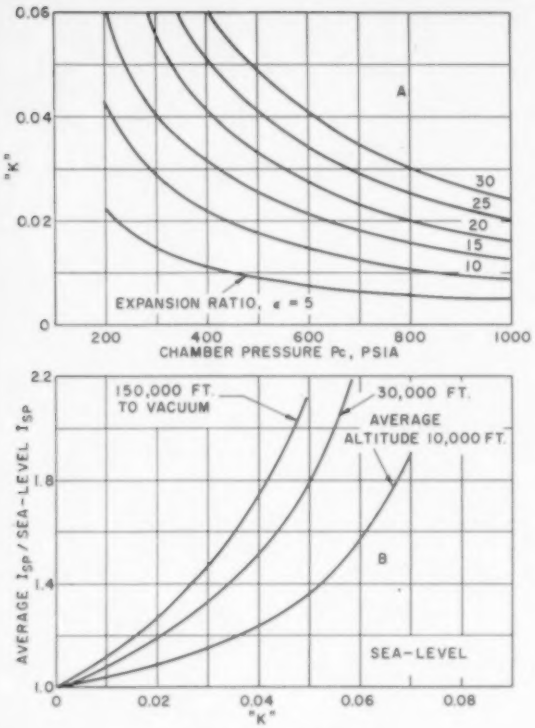


Fig. 1. Specific impulse ratio.

where $$K = \frac{e_x^{0.939}}{P_C}$$

Thus, Fig. 1*A* is a plot of K versus P_C for various values of e_x, expansion ratio. The plot in Fig. 1*B* was made from equation (6) using the appropriate value of P_{Avg}, the ambient pressure at the desired altitude. Since C_F is directly proportional to I_{SP}, the curve was plotted as a specific impulse ratio. A check of the resulting data indicated that Fig. 1 is valid for specific heat ratios between 1·05 and 1·40.

DRAG DATA (FIG. 2)

Drag data was obtained from the digital computer program, and by graphic integration the average drag was obtained as a function of mass ratio for different values of gravity turn "kick" angles. In order to evaluate the average drag, the missile area and average specific impulse are used in conjunction with Fig. 2.

MEAN ALTITUDE AND ALTITUDE MASS RATIO (FIGS. 3 AND 4)

In developing this method of trajectory

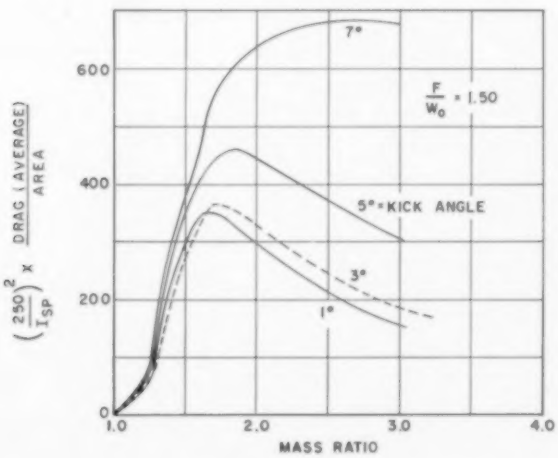


Fig. 2. Drag data.

and

$$\frac{C_{F\,Avg}}{C_{F\,S.L.}} = \frac{1{\cdot}485P_C - 1{\cdot}07e_x^{0.939}P_{Avg}}{1{\cdot}485P_C - 15{\cdot}72e_x^{0.939}}$$

$$= \frac{1}{1 - 10{\cdot}6K} - \frac{KP_{Avg}}{1{\cdot}386 - 14{\cdot}7K} \qquad (6)$$

analysis, it has been assumed that the effects of the atmosphere are negligible beyond 125,000 feet. Fig. 3 presents data on the mean altitude (between sea level and 125,000 ft) as a function of mass ratio at 125,000 ft for various kick angles. (Data was generated from digital computer

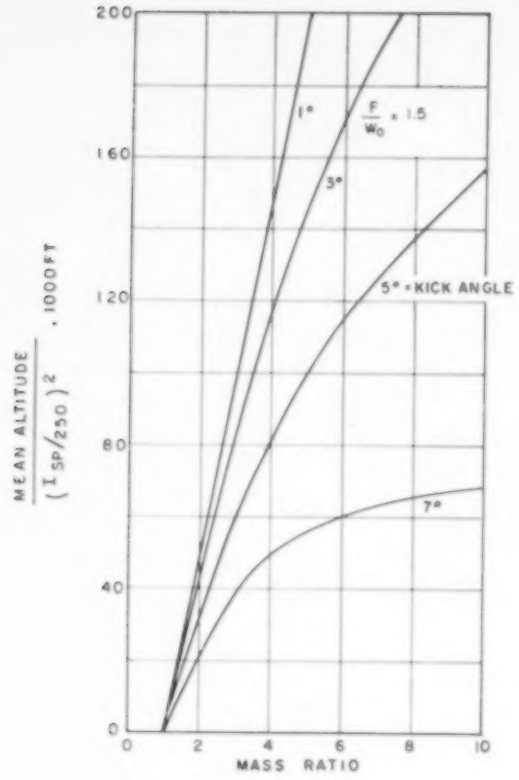


Fig. 3. Mean altitude data.

program with aid of graphic integration.) Fig. 4 was obtained by utilizing the data in Figs. 5 and 6. The curve presents the mass ratio at 125,000 ft altitude as a function of over-all

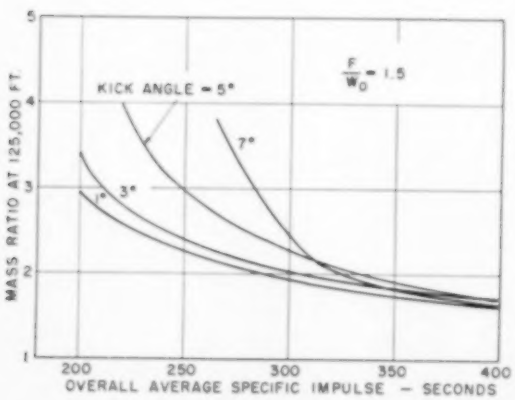


Fig. 4. Altitude mass ratio.

average specific impulse, i.e. the specific impulse corrected for the effects of drag and variation with altitude.

GRAVITY TURN VELOCITY, ALTITUDE AND RANGE (FIGS. 5-7)

Dimensionless plots of velocity and altitude are given in Figs. 5 and 6, respectively, for gravity turn flight paths. The data are presented

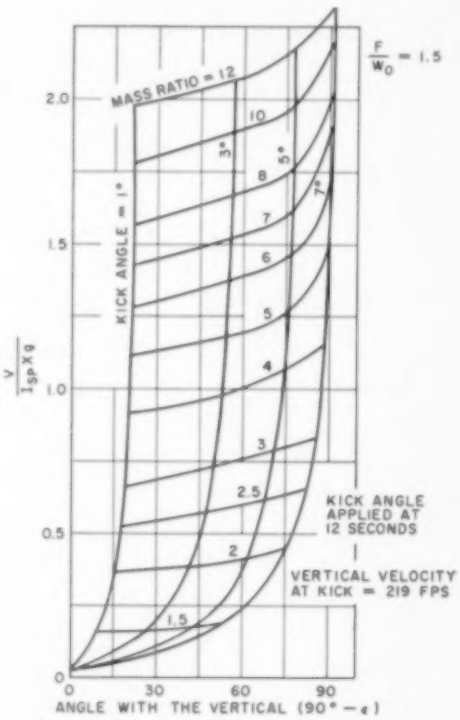


Fig. 5. Gravity turn velocity data.

as functions of the mass ratio and the angle of the velocity vector with the local vertical. Fig. 7 gives the range covered during powered flight.

CONSTANT ATTITUDE VELOCITY DATA (FIG. 8)

The data presented in this figure is for a constant attitude trajectory. For this type of flight path, the thrust vector angle relative to the local horizontal maintains a fixed value. It may

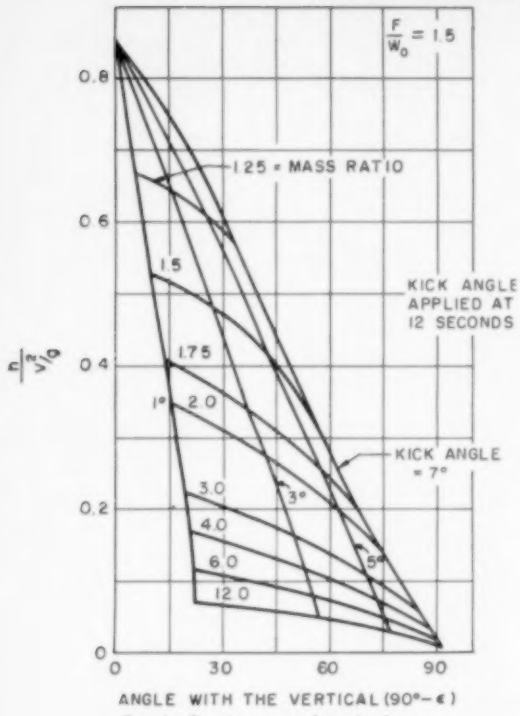


Fig. 6. Gravity turn altitude data.

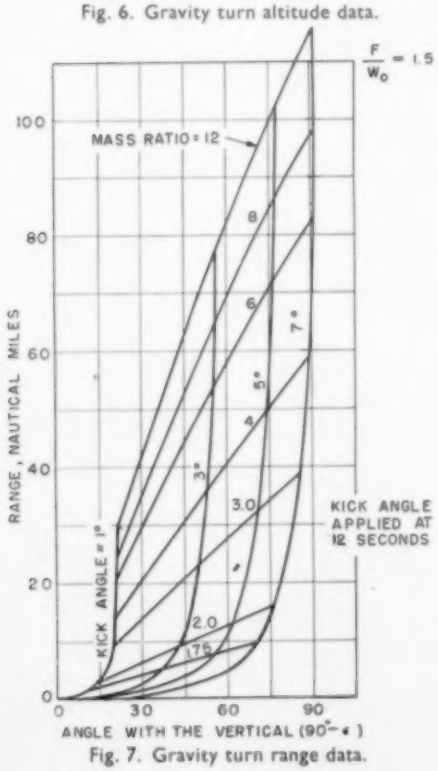


Fig. 7. Gravity turn range data.

be shown that this is a more efficient type of flight path, relative to energy expenditure. However, a gravity turn is usually utilized in the atmospheric portion of the flight to avoid the excessive angle of attack that might result with a constant attitude flight. It should be observed that Fig. 8 has an initial velocity as one of the input values. Curves for altitude and range, similar to Figs. 6 and 7, have been obtained but are not reproduced herein.

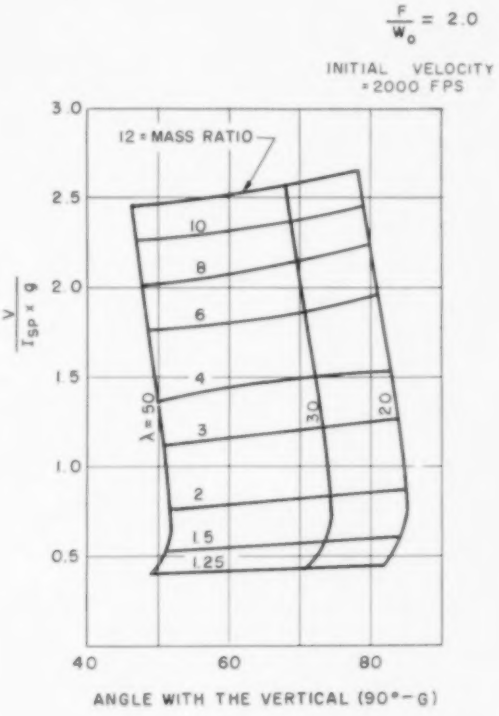


Fig. 8. Constant attitude velocity data.

RANGE CURVES (FIG. 9)

This figure shows range as a function of burnout angle, ϵ_{bo}, and velocity, V_{bo}. The range from this curve is the range covered during coasting and does not account for re-entry drag. The range covered during powered flight can be obtained from curves similar to Fig. 7. The total range during powered and coasting flight may then be calculated.

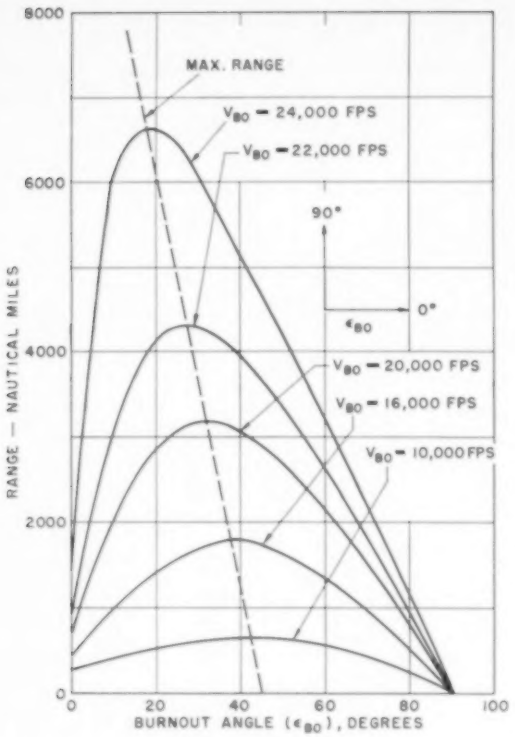


Fig. 9. Range data.

BALLISTIC MISSILE RANGE CALCULATION

This calculation is presented to illustrate a method for computing a ballistic missile range. For this, a gravity turn trajectory to obtain maximum range is desired. A single-stage missile has been assumed with the following data:

launch weight	100,000 lb
launch thrust	150,000 lb
burnout weight	12,500 lb
sea level specific impulse	250 sec
missile area	50 ft²
chamber pressure	600 psia
nozzle expansion ratio	15

The first step is to obtain the vehicle performance within the Earth's atmosphere (sea level to 125,000 ft altitude). The calculations will be given for a kick angle of 1°.

(A) $F/W_0 = 150{,}000/100{,}000 = 1{\cdot}5$.

(B) From Fig. 1*A*, using an expansion ratio of 15 and a chamber pressure of 600 psia, K is found to be 0·0213.

(C) Assume an average altitude, say 50,000 ft, for the atmospheric portion of the flight. Then, from Fig. 1*B*, $I_{SP_{Avg}}/I_{SP_{S.L.}} = 1{\cdot}24$.

(D) Average $I_{SP} = 1{\cdot}24 \times 250 = 310$ sec.

(E) Assume a value for average drag/average thrust say 0·05.

(F) Multiply item (D) by [1·0 − item (E)] to obtain over-all

$$I_{SP} = 310\,[1{\cdot}0 - 0{\cdot}05] = 295.$$

(G) From Fig. 4, using item (F) and the 1° kick angle curve, the mass ratio at 125,000 ft is equal to 1·97.

(H) From Fig. 2, at a mass ratio of 1·97, the drag factor is found to be 303.

(I) $$\text{Drag} = \text{item (H)} \times \text{area} \times \left(\frac{\text{item (F)}}{250}\right)^2$$

$$= 303 \times 50 \times \left(\frac{295}{250}\right)^2 = 21{,}100 \text{ lb.}$$

(J) $$\frac{\text{Average drag}}{\text{average thrust}} = \frac{\text{item (I)}}{F_{S.L.} \times \dfrac{\text{item (F)}}{I_{SP_{S.L.}}}}$$

$$= \frac{21{,}100}{150{,}000 \times \dfrac{295}{250}} = 0{\cdot}119$$

The value in item (J) should be compared with the assumption in item (E) and if reasonable agreement is not obtained, a new value should be chosen. A good approximation may be obtained by using the original assumption plus 75 per cent of the difference between this assumption and the value in item (J). Thus,

(E_1) $$\frac{\text{Average drag}}{\text{average thrust}}$$

$$= 0{\cdot}05 + 0{\cdot}75\,(0{\cdot}119 - 0{\cdot}05) = 0{\cdot}1016$$

(F_1) $310\,(1{\cdot}0 - 0{\cdot}1016) = 279$

(G_1) Mass ratio @ 125,000 ft = 2·05

(H_1) Drag factor = 288

(I_1) $\text{Drag} = 288 \times 50 \times \left(\frac{279}{250}\right)^2 = 17{,}850$

(J_1) $$\frac{\text{Average drag}}{\text{average thrust}} = \frac{17{,}850}{150{,}000 \times \frac{279}{250}} = 0{\cdot}1068$$

This checks sufficiently close with the original assumption, and the over-all average specific impulse between sea level and 125,000 ft is given as

$$310 \times (1{\cdot}0 - 0{\cdot}1068) = 277 \text{ sec}$$

(K) The assumption of average altitude may now be checked with the aid of Fig. 3, using the mass ratio of item (G_1) and a kick angle of 1°.

$$\text{Average altitude} = \left(\frac{\text{over-all } I_{SP}}{250}\right)^2 \times$$

$$\times\ 50{,}000 = \left(\frac{277}{250}\right)^2 \times 50{,}000 = 61{,}000 \text{ ft}$$

From Fig. 1*B* it may be observed that only a small difference in specific impulse ratio will result by the change in altitude from the assumed value of 50,000 ft in item (C). If a substantial difference were to result, the new altitude can be used in item (C) and the procedure repeated. Very rapid convergence should result.

(L) The velocity at 125,000 ft may be obtained with the aid of Fig. 5, using the mass ratio in item (G_1), the average specific impulse in item (J) and the kick angle of 1°. Thus,

$$V_{125{,}000 \text{ ft}} = \frac{V}{I_{SP} \times g}(I_{SP} \times g)$$

$$= 0{\cdot}386 \times 277 \times 32{\cdot}2 = 3440 \text{ ft/sec}$$

(M) The burnout mass ratio = 100,000/12,500 = 8·0.

Also, from Fig. 1*B*, the vacuum specific impulse may be found as

$$1{\cdot}29 \times 250 = 323 \text{ sec.}$$

The velocity attained during the remainder of the flight may then be computed with the aid of Fig. 3.

(N) $$V = I_{SP_{Vac}} \times 32{\cdot}2\left[\frac{V}{I_{SP} \times g}(\text{M.R.} = 8) - \right.$$

$$\left. - \frac{V}{I_{SP} \times g}(\text{M.R.} = 2{\cdot}05)\right]$$

$$= 323 \times 32{\cdot}2\ [1{\cdot}565 \times 0{\cdot}386] = 12{,}220 \text{ ft/sec}$$

The velocity at burnout, V_{bo}, is thus 3440 + 12,220 = 15,660 ft/sec. Also, from Fig. 3, the burnout angle with the vertical is found to be 21·5°, and from Fig. 5, the range covered during powered flight is given as 25 nautical miles. From Fig. 9, with a velocity of 15,660 ft/sec and an angle of 21·5°, the free flight range may be found as 990 nautical miles. Adding the range covered during powered flight, the result is 1015 nautical miles.

The above procedure may then be completed for kick angles of 3°, 5° and 7° and the results are as in Table 1.

Table 1

Kick angle (degrees)	Powered flight range (nautical miles)	Free flight range (nautical miles)	Total range (nautical miles)
1	25	990	1015
3	63	1950	2013
5	86	1360	1446
7	98	530	628

By plotting the range versus kick angle, the maximum range was found to be 2050 nautical miles for a kick angle of 2·75°.

An alternate means for flying this missile might utilize the gravity turn until an altitude of 125,000 ft was reached. At this time a constant attitude type of trajectory would commence with the thrust attitude angle being equal to the value of ϵ obtained at 125,000 ft. Utilizing curves similar to Fig. 8 and an initial velocity equal to that at 125,000 ft, the maximum range was found to be 2320 nautical miles, attained with a kick angle of 4°.

CONCLUSIONS

By use of generalized trajectory curves, a more accurate prediction of missile performance may be obtained than with the usual simplified procedures. Although the accuracy is somewhat less than that which can be realized with a digital computer, the results are well suited for preliminary design computations. The charts can be used for ballistic missile calculations and will aid in computing orbital trajectories.

Nomenclature

C_F = thrust coefficient
e_x = nozzle expansion ratio
F = thrust (lb)
f = factor to account for nozzle contraction ratio
g = gravitational constant at Earth's surface = 32·2 ft/sec²
h = altitude (ft)
I_{SP} = specific impulse (sec)
K = intermediate factor used to evaluate specific impulse ratio
M.R. = mass ratio
P = pressure (psi)
V = velocity (ft/sec)
W_0 = initial vehicle weight (lb)
ϵ = angle between velocity vector and local horizontal degrees
λ = angle between thrust direction and local horizontal (degrees)

Subscripts

A = ambient condition
Avg = average value
bo = value at burnout
C = value in combustion chamber
S.L. = value at sea level
Vac = value in vacuum

REFERENCE

1. R. C. Bowlin and K. J. Dresser, *One Dimensional Analysis of Rocket Motors*, Gen. Elect. Co. Rep. R57AGT511 (June 25, 1957).

A METHOD OF DESCRIBING MISS DISTANCES FOR LUNAR AND INTERPLANETARY TRAJECTORIES*

W. KIZNER
Jet Propulsion Laboratory, California Institute of Technology

Abstract—Miss distances for lunar and interplanetary trajectories can be described by specifying two components of the impact parameter, which is treated as a vector. This is analogous to the use of range and azimuth (or cross range) in specifying the impact point for terrestrial targets. The resulting co-ordinates are very nearly linear functions of the variables of the trajectory near the Earth, except in cases where the trajectory is of the minimum-energy type, such as a Hohmann orbit. Applications are given to the theory of guidance and to a method for automatically searching for a trajectory which hits or misses the target in a specified manner.

I. INTRODUCTION

In treating the errors of a vehicle trajectory which is intended to hit a target, such as Venus, or a point on the surface of the Earth, it is convenient to construct a system of differential corrections. The differential corrections are partial derivatives of some measure of the miss distance with respect to the co-ordinates at the burnout or injection point, and they are particularly important in the theory and design of guidance systems.

If the target is a point on the surface of the Earth, a satisfactory measure of miss distance is range and azimuth error. A typical differential correction, then, would be the partial derivative of range error with respect to speed at burnout.

For a "massy" point in space, however, some other measure of miss distance must be used. The seemingly obvious thing to do is to use the miss distance itself, i.e. the distance of closest approach of the vehicle to the center of the target. However if differential corrections are to be used for guidance or trajectory design purposes, then the use of this miss distance has certain drawbacks. In particular, for trajectories which pass close to the center of the target, the miss distance is a quadratic function of the initial conditions. Thus the variation of miss distance versus any of the initial co-ordinates is zero for a direct hitting trajectory. Another drawback in using the miss distance is that the direction of the path near the target is not given. For a terrestrial target this corresponds to an uncertainty in how much of the miss is due to an error in range, and how much to an error in azimuth.

A solution to the problem of specifying the miss distance is obtained by making use of the parameters of the hyperbola near the target. Components of the impact parameter, defined as in scattering theory in atomic physics but treated as a vector, are used to provide two quantities which eliminate the nonlinearities caused by the gravitational field of the target, and specify the direction of the miss.

The general problem of how many independent variables are observable near the target is investigated. This leads to a discussion of guidance equations.

II. A MEASURE OF MISS DISTANCE

The specification of miss distance that is used is based on the impact parameter, **B**, which is computed from the osculating conic (see appendix). The elements of the conic are computed when a fixed distance from the target is reached, the distance being determined by the point where a conic approximation for the remainder of the trajectory is sufficiently accurate,

* This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract No. NASw-6, sponsored by the National Aeronautics and Space Administration.

or when closest approach is attained, whichever occurs first. This is further discussed in the appendix. **B**, the position vector in the plane of the trajectory originating at the center of gravity of the target and directed perpendicularly to the incoming asymptote of the hyperbola (see Fig. 1), is approximately the vector miss which would occur if the target had no mass.

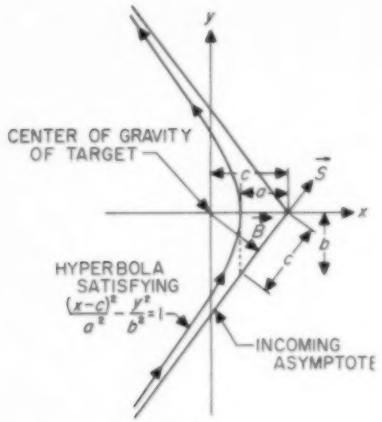


Fig. 1. Geometry near target.

In constructing a system of differential corrections, all three components of **B** cannot be used because they are not independent. Since the asymptote of the hyperbola is approximately fixed in direction (for interplanetary trajectories it varies less than 0·01° for expected changes in initial conditions), the components of **B** must satisfy the relation that **B** is orthogonal to the asymptote.

To account for this, two components of **B**, which are analogous to the two quantities used to describe the miss for a terrestrial target, are used. A set of orthogonal unit vectors is computed. Let **S** be a unit vector in the direction of the incoming asymptote, **T** be a unit vector perpendicular to **S** that lies in a fixed plane such as the equatorial plane, and **R** be a unit vector to form a right-handed system.

$$\mathbf{R} = \mathbf{S} \times \mathbf{T} \tag{1}$$

Since **B** is perpendicular to **S**, it lies in the plane determined by **R** and **T**. The variables which are used are the projections of **B** on **T** and on **R**, i.e. $\mathbf{B}\cdot\mathbf{T}$ and $\mathbf{B}\cdot\mathbf{R}$. The complete set of formulas for computing the vectors is given in the appendix.

The relation between the miss distance and **B** will now be shown. If a particular value of the distance of closest approach is desired, the magnitude of **B** can be computed from the following relations concerning the parameters of the hyperbola. In rectangular co-ordinates

$$\frac{(x-c)^2}{a^2}-\frac{y^2}{b^2}=1 \tag{2}$$

where a is the semimajor axis,

$$c=\sqrt{(a^2+b^2)} \tag{3}$$

$$b=|\mathbf{B}| \tag{4}$$

a is essentially determined by the launch time and the time of flight, hence it can be determined from a trajectory which does not have the desired miss distance. Since the miss distance is $c-a$, b can be determined. For interplanetary trajectories, a can be neglected in comparison to b for large miss distances, such as 50 target diameters, so that the miss distance is then approximately equal to b. By the components of **B**, or $\mathbf{B}\cdot\mathbf{T}$ and $\mathbf{B}\cdot\mathbf{R}$, the direction of the miss is determined.

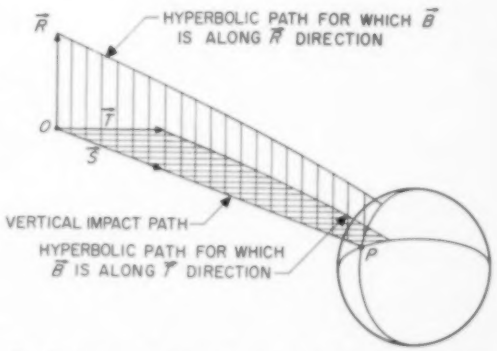


Fig. 2. Location of impact points on the surface of the target.

These relations are shown in Fig. 2. The line OP represents the path directed at the center of the target. This line also represents the asymptote of the osculating conic, which is very nearly

parallel to the asymptotes of other conics which intersect the surface of the target. Consider now the cases where the impact is not vertical and the impact parameter is along the **T** axis. Since the plane of the trajectory contains the original **S** vector the locus of the impact points on the surface of the target must be a great circle. This is true whenever the direction of **B** is held fixed and the magnitude is varied. It can also be seen that the angle between such great circles at P is the same as the angle between the corresponding **B** vectors.

The orientation of the asymptote has been investigated for many different kinds of trajectories. The greatest variations in the orientation occur in lunar trajectories, but even here the direction is fixed to within a fraction of a degree for expected variations in initial conditions.

III. APPLICATIONS TO HOMING

In searching for trajectories which miss the target in a specified manner, one often has approximate values of the parameters that determine the powered flight or the values of the co-ordinates at the end of the powered flight. Better estimates of the initial conditions are needed to insure a specified miss distance. The method used in accomplishing this is called homing.

The homing method proceeds by computing a reference trajectory based on the original estimate of the initial conditions, and then computing two trajectories where each one of two initial co-ordinates or parameters is varied while the other is kept fixed. By observing the resulting changes in $\mathbf{B}\cdot\mathbf{T}$ and $\mathbf{B}\cdot\mathbf{R}$ due to small changes in the input conditions, new input conditions can be obtained by a linear interpolation scheme. If nonlinearities are suspected, then the size of the increments to the initial variables can be monitored so that they do not exceed values for which the linear interpolation method succeeds. Table 1 shows the number of iterations required by the homing method for two Venus trajectories. The time of launch and the firing angle were varied here.

In Table 1 the perihelion distance is the minimum distance to the Sun of the transfer ellipse. It can be seen that in Trajectory B, where the perihelion distance is smaller, the number of trials required is less. Whenever the perihelion distance of the transfer conic is nearly equal to the distance of Venus from the Sun then the differential corrections become nonlinear and more iterations are required in homing.

Table 1. Example of the Homing Method for Two Venus Trajectories.

Parameter		Trajectory A	Trajectory B
Time of flight (days)		107	96
Perihelion distance (km)		$1{\cdot}02 \times 10^8$	$9{\cdot}5 \times 10^7$
Distance to Venus from the Sun (km)		$1{\cdot}08 \times 10^8$	$1{\cdot}08 \times 10^8$
Miss distance (km)	Trial 1	$3{\cdot}1 \times 10^6$	$1{\cdot}6 \times 10^6$
	Trial 2	$9{\cdot}6 \times 10^5$	$5{\cdot}0 \times 10^4$
	Trial 3	$2{\cdot}8 \times 10^5$	3·2
	Trial 4	$1{\cdot}4 \times 10^3$	
	Trial 5	0·13	

The kind of lunar trajectories for which the homing method has been used is not the minimum-energy type. Generally only one or two iterations have been required to achieve the correct miss distance.

An interesting case occurred in homing on a Mars trajectory when a mistake was made in the input conditions and the right ascension was in error by 180°. Nevertheless, the homing method succeeded in varying the initial conditions such that a satisfactory miss distance could be attained.

IV. DEGREES OF FREEDOM

In order to generalize the ideas about miss distance and to introduce problems in guidance the following question will be treated: Given a launching location on the Earth's surface, a vehicle configuration and a range of firing times (limited by the rotation of the Earth), what is the minimum number of quantities that can be specified concerning the trajectory near the target in order that the trajectory shall be completely determined? The answer to this question is connected with the formation of

guidance equations, which must ensure that all or some of these quantities near the target attain prescribed values.

If the trajectory were not constrained to originate from a certain region at the Earth and start during a given time interval, there would be six quantities that could be specified about the trajectory near the target corresponding to the six degrees of freedom of the vehicle. With the constraints imposed by the launching conditions there are only three degrees of freedom. This was discovered by examining many lunar trajectories and finding the subspace which contains the points of the allowable trajectories.

The reason for the loss of half of the degrees of freedom can be found by considering an interplanetary trajectory, such as one which approaches Mars. Instead of considering the actual trajectory, consider an equivalent trajectory where the path in the Sun's region of influence is the same, but where the path near the Earth is replaced by a fictitious path where the vehicle has an initial velocity near the Earth and the Earth is assumed to have no effect on the vehicle. In other words, the path near the Earth is approximated by the asymptote to the original escape hyperbola. The family of asymptotes which represent the class of trajectories approaching Mars then seems to originate from a small region, within an order of magnitude of the size of Earth. Since this size is insignificant compared to the distances involved, the equivalent trajectories may be thought of as originating from a point source. There are three velocity components that can be varied. This corresponds to the three degrees of freedom near the target.

The essential part of the proof is that the trajectories seem to originate from a point source. A simplified version of the problem may be considered in which all gravitational fields are neglected. Trajectories then are straight lines which originate in one region in space and are directed towards a target point. It can easily be proved that as the distance between the target and the region of origin is increased, only three numbers are necessary to specify the trajectory near the target.

Two co-ordinates, $\mathbf{B}\cdot\mathbf{T}$ and $\mathbf{B}\cdot\mathbf{R}$, have been described; it remains to choose a third co-ordinate. The time of flight has been used as the third co-ordinate since this is often of direct interest. For a mission such as a soft landing on the moon the *vis viva*[1] or total energy near the target might be chosen, since the velocity at the surface of the target is then readily available. The *vis viva* and also the semimajor axis a are more linear functions than the time of flight, and therefore either one is an excellent choice for a third variable.

If this particular co-ordinate system is linear with respect to initial co-ordinate changes for a given mission, then since the system is complete (or has the maximum number of dimensions) any other system of co-ordinates which is linear must be a linear transformation of this system. Thus for missions which possess the linearity property, the given co-ordinate system can be used as a basis for any other system which is more convenient.

V. GUIDANCE EQUATIONS

Having discussed the meaning of the set of co-ordinates, their differential corrections will now be discussed. It will be shown that differential corrections can lead to guidance equations. In particular the use of a set of co-ordinates which is fairly linear provides some insight as to what form the guidance equations may take.

A method of analysing guidance equations is to think of the equations as predicting the co-ordinates near the target. Let Q_i ($i = 1, 2$ and 3) be the co-ordinates near the target, and q_j ($j = 1, 2, \ldots 6$) be the co-ordinates in the region where the vehicle is being guided. For simplicity assume that only linear terms need be used. Then the guidance equations are

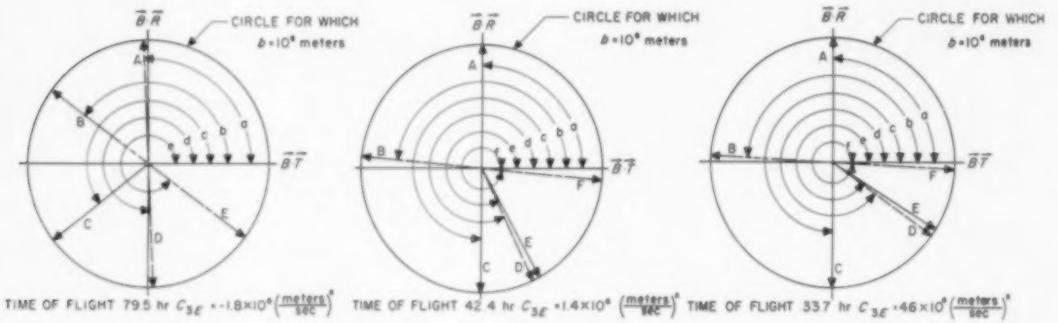
$$\delta Q_i = \sum_{j=1}^{6} \frac{\partial Q_i}{\partial q_j}\,\delta q_j \quad i = 1, 2, 3 \qquad (5)$$

The variations δq_j refer to instantaneous variations from the standard value. By steering the vehicle or by shutting the motor off at a non-standard time the δQ_i can be nulled. It should be

noted that the guidance equations incorporate the differential corrections.

Not all of these equations need be used. For instance if the time of flight need not be controlled, then one equation is eliminated. If the standard point of closest approach to the target is far removed from the center of the target it may be feasible to eliminate another constraint

In this set of variables a change in θ is equivalent to a change in the ground range. The effects of changes in these variables for lunar trajectories are illustrated in Fig. 3. Here the values of the co-ordinate deviations for the initial co-ordinates were chosen so that $|\mathbf{B}|$ is equal to 1000 km. The trajectories chosen all have the angle between the orbital plane near



CASE A. $\Delta Z = 0.185017\times10^8$ meters
a. 323.9 deg
CASE B. $\Delta\theta = 0.191702$ deg
b. 2. 272 deg
CASE C. $\Delta R = 0.718876\times10^8$ meters, $\Delta V = 0.601450$ meters/sec
c. 218 deg
CASE D. $\Delta\dot{Z} = 114.103$ meters/sec
d. 143.5 deg
CASE E. $\Delta\Gamma = 0.949780\times10^{-1}$ deg
e. 91.2 deg

CASE A. $\Delta Z = 0.189986\times10^8$ meters
a. 90 deg
CASE B. $\Delta\theta = 0.155277$ deg
b. 174.5 deg
CASE C. $\Delta\dot{Z} = 59.1199$ meters/sec
c. 270 deg
CASE D. $\Delta R = 0.598090\times10^8$ meters
d. 296.5 deg
CASE E. $\Delta V = 5.09664$ meters/sec
e. 299 deg
CASE F. $\Delta\Gamma = 0.798537\times10^{-1}$ deg
f. 354.5 deg

CASE A. $\Delta Z = 0.201213\times10^8$ meters
a. 90 deg
CASE B. $\Delta\theta = 0.152333$ deg
b. 178 deg
CASE C. $\Delta\dot{Z} = 49.2961$ meters/sec
c. 4. 270 deg
CASE D. $\Delta V = 0.677462\times10$ meters/sec
d. 325 deg
CASE E. $\Delta R = 0.763812\times10^8$ meters
e. 328 deg
CASE F. $\Delta\Gamma = 0.806707\times10^{-1}$ deg
f. 357 deg

Fig. 3. Effect of co-ordinate error for lunar trajectory.

and to guide along the direction from the standard miss distance to the center of the target and to disregard errors in the perpendicular direction. A useful set of variables to use to describe errors in the guided portion of the trajectory are

- R = the distance from the center of the Earth (or other body)
- V = speed
- θ = angle between the standard perigee direction and the radius vector to the vehicle
- Z = cross range (which is zero on the standard path)
- $\dot{Z}$ = cross range rate
- Γ = path angle measured with respect to the local horizontal.

the Earth inclined by an angle of 30° to the plane of the Moon, and have an angle of 10° between the point of injection and perigee. Three different values of the *vis viva* near the Earth, C_{3E}, were used.

It can be seen that in the three cases shown, and in practically all the cases that have been computed which represent all reasonable lunar trajectories, that the co-ordinates R and V, θ and Γ, Z and $\dot{Z}$ can be paired insofar as the effects of the miss near the Moon are concerned. For example an error in ground range θ can be compensated for by a change in path angle Γ. Another feature which emerges is that the effect of a change in Z is opposite to that produced by a change in $\dot{Z}$. Hence a scheme to cancel the effects of one by introducing a change

in the other is unstable from the point of view of incurring large deviations from the standard flight path.

In Fig. 4 a similar plot is shown for a Mars trajectory. It shows a similarity to a fast lunar trajectory.

compute the elements of the conic when the vehicle has not yet reached the point of closest approach. For lunar trajectories in which the aspherical nature of the Moon is neglected, it was found that a good point to compute the conic is when the probe is about three lunar

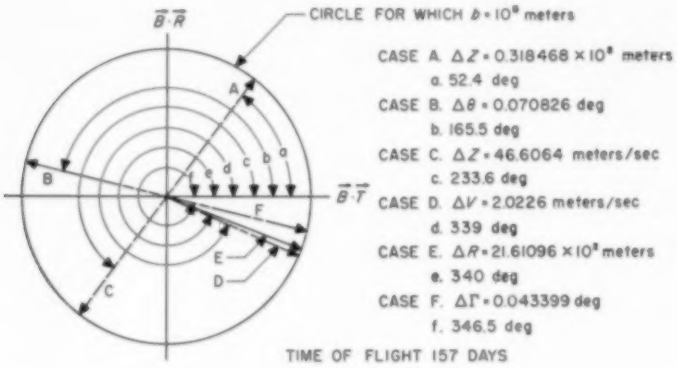


Fig. 4. Effect of co-ordinate error for Mars trajectory.

VI. CONCLUSIONS

It has been shown that the variables $\mathbf{B}\cdot\mathbf{T}$ and $\mathbf{B}\cdot\mathbf{R}$ can serve as a measure of miss distance which removes the nonlinearity caused by the attraction of the target. The variables are a function of the osculating conic near the target and as such can be exactly computed for any lunar or interplanetary trajectory. With the addition of a third variable, such as the time of flight, the set of variables near the target is complete. The set of variables can be used in much the same manner as range and azimuth error in analysing terrestrial trajectories.

APPENDIX

The Calculation of Elements of the Osculating Conic near the Target

The point at which the osculating conic near the target is computed will depend on whether the method of integration is direct (such as Cowell's method) or based on a variational scheme (for instance Encke's method). For a direct method the accuracy of the computation decreases when the path is near the attracting center, so that it may be advantageous to compute the elements of the conic when the probe is about three lunar diameters from the center of the Moon. So that at this point or at the point of closest approach, whichever occurs first, the elements of the conic are computed. The method of computation is discussed in Herrick.[2] It is assumed that the rectangular co-ordinates $(x,y,z,\dot{x},\dot{y},\dot{z})$ of the relative motion of the vehicle to the center of mass are available.

$$r = \sqrt{(x^2 + y^2 + z^2)} \tag{6}$$

$$\dot{s} = \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)} \tag{7}$$

$$r\dot{r} = x\dot{x} + y\dot{y} + z\dot{z} \tag{8}$$

$$r^4\dot{v}^2 = r^2\dot{s}^2 - (r\dot{r})^2 \tag{9}$$

The *semilatus rectum*,

$$p = \frac{r^4\dot{v}^2}{GM} \tag{10}$$

where G is the universal gravitational constant and M is the mass of the target.

$$\frac{1}{a} = \frac{2}{r} - \frac{\dot{s}^2}{GM} \tag{11}$$

The square of the eccentricity,

$$e^2 = 1 - \frac{p}{a} \tag{12}$$

The angular momentum constant,

$$c_1 = r^2\dot{v} \tag{13}$$

The *vis viva*,

$$c_3 = -\frac{GM}{a} \tag{14}$$

The sine and cosine of the true anomaly are given by

$$\cos v = \frac{1}{e}\left(\frac{p}{r} - 1\right) \tag{15}$$

$$\sin v = \frac{1}{e}\dot{r}\sqrt{\left(\frac{p}{GM}\right)} \tag{16}$$

The unit vector along the radius direction

$$\mathbf{U} = \frac{\mathbf{r}}{r} \tag{17}$$

An auxiliary vector $\mathbf{V}$ is found from

$$c_1\mathbf{V} = r\dot{\mathbf{r}} - \dot{r}\mathbf{r} \tag{18}$$

The vector normal to the orbital plane

$$\mathbf{W} = \frac{1}{c_1}\mathbf{r} \times \dot{\mathbf{r}} \tag{19}$$

The vector directed towards pericenter

$$\mathbf{P} = \mathbf{U}\cos v - \mathbf{V}\sin v \tag{20}$$

To complete the system

$$\mathbf{Q} = \mathbf{U}\sin v + \mathbf{V}\cos v \tag{21}$$

is defined.

The impact parameter magnitude is found from

$$b = \frac{c_1}{\sqrt{|c_3|}} \tag{22}$$

The unit vector in the direction of the incoming asymptote

$$\mathbf{S} = -\frac{a}{\sqrt{(a^2 + b^2)}}\mathbf{P} + \frac{b}{\sqrt{(a^2 + b^2)}}\mathbf{Q} \tag{23}$$

$$\mathbf{B} = \frac{b^2}{\sqrt{(a^2 + b^2)}}\mathbf{P} + \frac{ab}{\sqrt{(a^2 + b^2)}}\mathbf{Q} \tag{24}$$

In the last two expressions a is negative.

Also

$$T_x = \frac{S_y}{\sqrt{(S_x^2 + S_y^2)}} \tag{25}$$

$$T_y = -\frac{S_x}{\sqrt{(S_x^2 + S_y^2)}} \tag{26}$$

$$T_z = 0 \tag{27}$$

and

$$\mathbf{R} = \mathbf{S} \times \mathbf{T} \tag{28}$$

The reader may prefer to define $\mathbf{R}$ by the negative of this expression to make a positive $\mathbf{R}$ correspond to what is normally considered up with respect to the equatorial or ecliptic planes. Also, a direct (as opposed to a retrograde) path corresponds to a positive $\mathbf{T}$.

REFERENCES

1. F. R. Moulton, *Introduction to Celestial Mechanics*, MacMillan, New York (1958).
2. D. Herrick, *Astrodynamics*, Van Nostrand, in press.

Some General References

3. R. H. Battin and J. H. Laning, Jr., *A Navigation Theory for Round Trip Reconnaissance Missions to Venus and Mars*, Massachusetts Institute of Technology, Cambridge, Instrumentation Laboratory (Aug. 1959).
4. A. B. Mickelwait and R. C. Booton, Jr., *Analytical and Numerical Studies of Three Dimensional Trajectories to the Moon*, presented at the Institute of the Aeronautical Sciences, Los Angeles, California, June 16–19, 1959.
5. A. B. Mickelwait, E. H. Tomkins, Jr., and Parks, R. A., *Three Dimensional Interplanetary Trajectories*, Trans. Inst. Radio Engrs. on Mil. Electronics, p. 119 **3**, (1959).

ORBITAL TRANSFER FOR SATELLITES

SAMUEL P. ALTMAN
The Martin Company, P.O. Box 179, Denver 1, Colorado

Abstract—Orbital transfer between elliptical orbits is formulated by six integral equations with six orbital parameters. The planar orbit is defined by two velocity parameters rather than the conic parameters. Consequently, the orbital and transfer velocities and accelerations are presentable as hodographs of radial and normal components. The set of integral equations may be separated for analysis into two sets of three equations each whenever certain conditions are fulfilled.

Planar orbital transfer by use of two velocity increments is presented in a general formulation, of which the Hohmann transfer ellipse is a special case.

INTRODUCTION

The development of artificial satellites requires the attainment of specified orbits, to assure successful completion of the desired missions. The objectives of such missions may be research, military or exploratory. In any case, terrestrial reconnaissance, astrophysical observations, departure conditions for interplanetary flight and rendezvous for satellite maintenance or crew changes will require control and correction of a satellite orbit or the trajectory of an orbital transfer vehicle.

The correction of a satellite orbit may be necessary when the desired final orbit cannot be provided by available launch- and ascent-guidance systems. Then low-level (or vernier) thrust application may provide reorientation and reshaping of the orbit.

An orbital transfer vehicle intended for satellite maintenance or crew changes will probably be established initially in orbit in close proximity to the satellite, following the ascent trajectory phase of flight. Then the position and velocity of the transfer vehicle and satellite must be matched so that the two bodies may be joined in a common (or mutually fixed) orbital system without shock and damage, and without primary disturbance of the satellite orbit. Although both the transfer vehicle and the satellite might be maneuvered by thrust application, most foreseeable missions will probably require that the satellite remain in a given reference orbit. Not only will the mission objectives be so fulfilled, but the predictions of orbit derived from prior ground-tracking observation and computation will then be valid.

Orbital transfer of a vehicle to join (or rendezvous with) a satellite implies an added restriction upon the solution, which is not present in the problem of orbital transfer for a satellite alone. Namely, the respective positions (or phases) in orbit at a given time must be further considered.

The selection of a transfer trajectory between orbits is characterized by the typical analytical difficulties associated with most space flight problems. Out of an infinite number of possible trajectories, those trajectories that fulfill all end-point conditions must be determined. Finally, the classes of trajectories that provide at least partial optimalization of critical performance factors, and the definitive characteristics of those trajectories must be deduced. The nonlinear equations of motion do not facilitate simple, direct analysis of the transfer problem.

In general, an orbit about a spherical celestial body will be closed (or periodic) and planar, while an orbit about a spheroidal body will be conditionally periodic. For example, an orbit about an oblate spheroid (which closely approximates the terrestrial body) will appear as shown schematically in Fig. 1. This conditionally periodic orbit can be defined analytically (see Fig. 2) by an elliptical closed orbit in a plane that rotates about the minor (or polar) axis. The line of intersection (or line of nodes) of the orbital plane with the equatorial plane is defined by Ω, measured in the equatorial plane from a given equatorial reference axis; the

included angle (or angle of incidence) between planes is defined by i. The radial projection of the perigee (or apsidal point of closest approach of the orbital vehicle to the celestial body) upon the unit sphere is defined by the apsidal ψ, Ω, a, e and T define the orbit in classical celestial mechanics.

Brouwer, [1]Orlov[2] and Blitzer *et al.*[3, 4] demonstrated that the oblateness of the celestial spheroid will cause rotation of the line of nodes

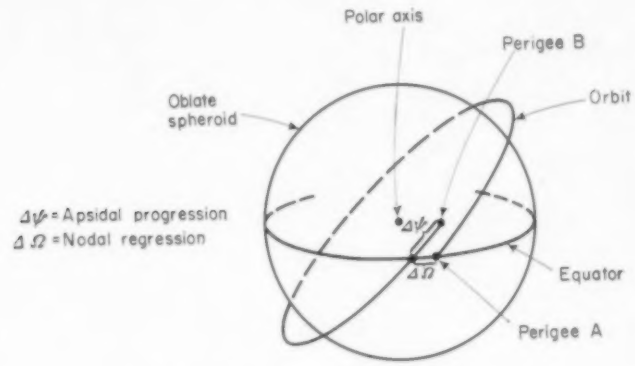


Fig. 1. Conditionally periodic orbit.

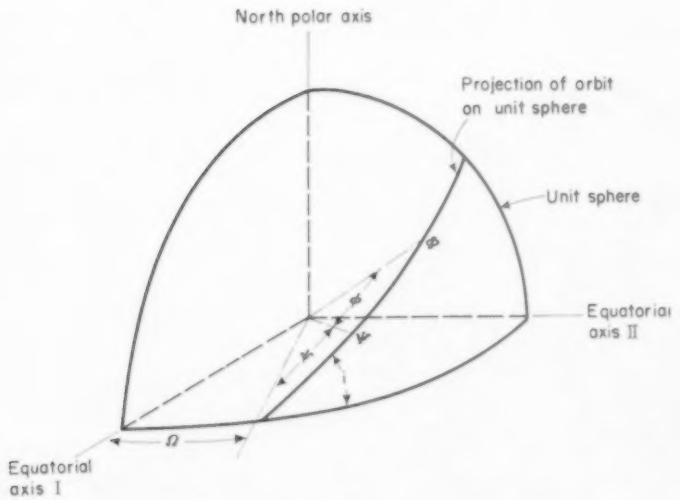


Fig. 2. Definition of orbital plane orientation.

line Ψ. The angular displacement of the perigee from the line of nodes is measured by ψ. The shape of the elliptical orbit is defined by the conic parameters a (the semi-major axis) and e (the eccentricity). Finally, the time T of perigee passage of the vehicle in orbit enables definition of its phase angle ϕ (or true anomaly) in its orbit. In summary, the orbital elements i, for the orbital plane and the line of apsides for the elliptical orbit. Brouwer analysed nearly circular orbits in the vicinity of the equatorial plane, Orlov studied nearly circular orbits in a plane of arbitrary (or any) incidence angle, and Blitzer and his associates investigated the general cases of circular and elliptical orbits in a plane of arbitrary incidence angle. While correlation

exists between these respective works, the most general results due to Blitzer *et al.* show that the nodal and apsidal rotations are expressed as functions of the orbital elements, respectively as follows:

$$-\dot{\Omega}_J = \frac{JR_e^2 \cos i}{a^2(1-e^2)^2} n \tag{1}$$

$$\dot{\psi}_J = \frac{JR_e^2 (2 - \frac{5}{2}\sin^2 i)}{a^2(1-e^2)^2} n \tag{2}$$

These expressions were derived by consideration only of the coupling constant J of the spherical harmonic expansion for the gravitational potential field of an oblate spheroid. The coupling constant D and higher order terms are negligible in the orbital transfer problem, compared with the coupling constant J.

A transfer trajectory between two orbits will be provided by thrust application to the transfer vehicle, so that cross-plane velocity and in-plane velocity change will occur. A cross-plane velocity due to thrust will change i, Ω and ψ, while an in-plane velocity change due to thrust will change ψ, a, e and T.

Since an orbit is most effectively expressed in terms of six orbital parameters, a transfer trajectory will be defined by a set of six integrals which are integrated over the time interval of transfer. Since acceptable transfer is only provided by matching both position and velocity, the integrands are necessarily velocity terms. However, the conic parameters a and e are not amenable to simple mathematical statement and manipulation. Upon study of the total energy equation, it became apparent that an orbit can be simply defined by two velocity parameters, rather than the conic parameters. Consequently, the transfer trajectory will be defined by six integrals of the two velocity parameters and the time-derivatives for i, Ω, ψ and ϕ. Orbit reshaping for a satellite will render the phase-integral superfluous, so that an orbit reshaping trajectory will be defined by five integrals.

This formulation of the orbital transfer problem presents the necessary and sufficient conditions for acceptable transfer, in complete analytic form. Although the nature of the transfer trajectory itself must still be selected, this formulation enables such selection to be predicated on achievable performance characteristics. For example, the accelerations defining the planar figure of the orbit and the transfer accelerations are derived from the velocity parameters. Both velocity and acceleration of either orbit or transfer trajectory can be presented graphically on a polar hodograph. Consequently, the time-history of thrust magnitude and direction may be directly related to the velocity parameters, so that the engineering design of propulsion and staging may be more easily assessed in meeting the trajectory design requirements for orbital transfer.

Study of this transfer formulation enables the simple derivation of those trajectory design conditions which—if fulfilled—will permit separation of the complete problem into the simpler cases of cross-plane velocity defining incidence, nodal and apsidal variations, and in-plane velocity defining the velocity parameter and phase changes. Also, the oblateness effect may be related to the transfer problem, so that error in solution due to its omission can be readily evaluated.

Finally, the many arbitrarily selected trajectories, which might fulfill the conditions for acceptable transfer, may be evaluated by the additional use of auxiliary conditions for optimalization (e.g. time, fuel). This formulation appears amenable to future optimalization study. Transfer in an orbital plane by use of velocity increments is developed by use of the polar hodograph. The Hohmann transfer ellipse[5] is discussed and shown to be a special case of the more general class of two-impulse planar transfer.

REPRESENTATION OF THE PLANAR ORBIT BY THE POLAR HODOGRAPH

The total potential and kinetic energy of orbital motion for a satellite in a conservative energy state, and in the gravitational field of a celestial sphere can be presented in the following form:

$$E = \frac{1}{2} m(v_\phi^2 + v_r^2) - \frac{\kappa}{r} \tag{3}$$

Since the angular momentum of a conservative orbital system is constant (i.e. equal to h), then

$$r = \frac{h}{mv_\phi} . \tag{4}$$

Consequently,

$$\frac{2E}{m} = v_r^2 - \left(\frac{2\kappa}{h}\right) v_\phi + v_\phi^2 \tag{3a}$$

so that

$$\frac{2E}{m} + \left(\frac{\kappa}{h}\right)^2 = v_r^2 + \left(v_\phi - \frac{\kappa}{h}\right)^2 . \tag{3b}$$

Equation (3b) is seen to represent a circle, when presented in the form

$$R^2 = v_r^2 + (v_\phi - C)^2 \tag{3c}$$

where

$$C = \frac{\kappa}{h} \tag{5}$$

and

$$R = \sqrt{\left[\frac{2E}{m} + \left(\frac{\kappa}{h}\right)^2\right]} = \sqrt{\left[\frac{2E}{m} + C^2\right]} . \tag{6}$$

This circle, shown schematically in Fig. 3, defines the orbit uniquely in terms of the radial and normal velocities. This circle is the polar hodograph of the orbit, so named since it presents the polar co-ordinate components of the velocity vector producing the classical hodograph. Although the classical hodograph is a circle that defines the magnitude and direction of the velocity vector directly, it does not permit the presentation of two or more orbits in a common reference system.

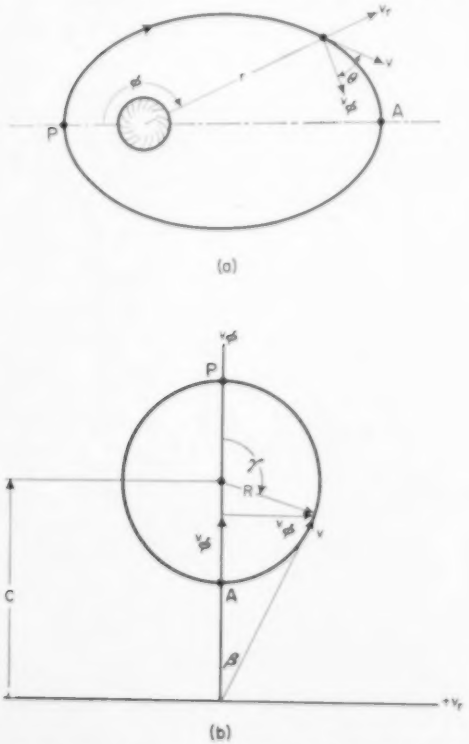


Fig. 3. Planar orbit in space (a) and its hodograph (b).

The rectilinear reference system of normal and radial velocities (v_ϕ, v_r) provides a common reference for all orbits about the one celestial body. However, the v_ϕ axis defines the line of apsides for a given orbit, although the lines of apsides for two or more orbits may not coincide in space. The magnitude of the total velocity is obtainable directly from the polar hodograph. Further study of the polar hodograph and equation (4) shows that the exterior intercept of the hodograph circle and the v_ϕ axis defines perigee, while the interior intercept defines apogee. The following relations which map the orbit in space into the polar hodograph plane have been derived by analysis:

$$\phi = \gamma \tag{7}$$

$$\theta = \beta \tag{8}$$

$$e = \frac{R}{C} \tag{9}$$

$$a = \frac{\mu}{C^2 - R^2} \tag{10}$$

$$n = \frac{(C^2 - R^2)^{3/2}}{\mu} \tag{11}$$

$$a(1 - e^2) = \frac{\mu}{C^2} . \tag{12}$$

A circular orbit ($R = 0$) becomes a point on the ordinate (v_ϕ) axis at a constant distance C from the origin.

The velocity of the vehicle in orbit is determined at any point by the relations

$$v_r = R \sin \phi , \tag{13}$$

$$v_\phi = C + R \cos \phi \tag{14}$$

and

$$v_r^2 + (v_\phi - C)^2 = R^2 . \tag{15}$$

The subsequent use of these relations is governed by the condition that the true anomaly ϕ is referred to perigee; that is,

$$\phi = \sigma - \psi \,. \tag{16}$$

The graphical relations between the hodographs of two orbits in space are shown schematically in Fig. 4. The hodographs are separate only if the perigee of the larger orbit is greater than the apogee of the smaller orbit. The hodo-

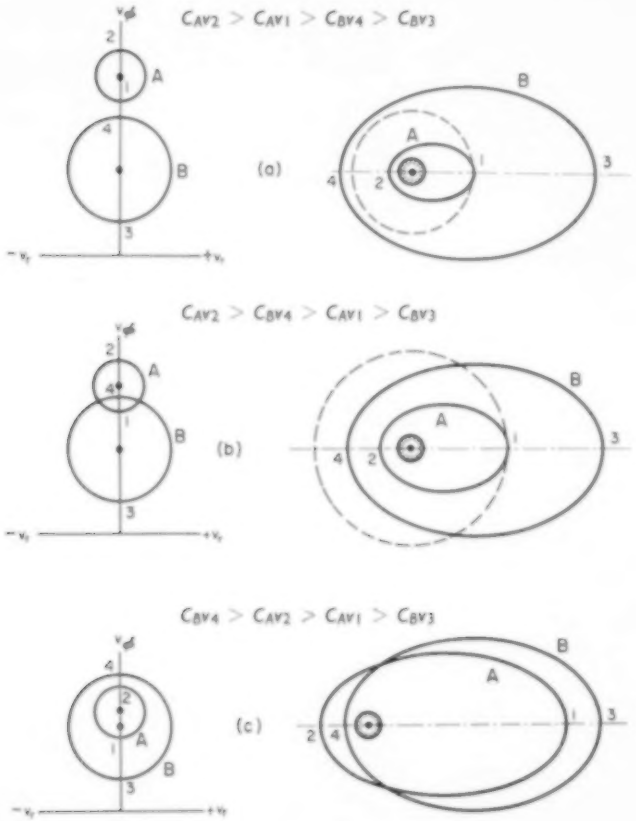


Fig. 4. Graphical relations between orbital hodographs:
(a) orbits with isolated hodographs; (b) orbits with intersecting hodographs;
(c) orbits with circumscribing hodographs.

The velocity parameters (C, R) may be presented in the following forms:

$$C = \frac{\mu}{r v_\phi} \tag{5a}$$

$$R = \sqrt{(2\bar{E} + C^2)} \,. \tag{6a}$$

Since $\bar{E}$ is the specific total orbital energy per unit mass, both velocity parameters are independent of the system mass. However, the R parameter is a function of the C parameter.

graph of one orbit circumscribes the hodograph of a second orbit if $v_4 > v_2 > v_1 > v_3$ as shown in Fig. 4c.

ORIENTATION OF THE ORBITAL PLANE

The orbital plane is defined in inertial space by the center of the attracting celestial sphere and the total velocity vector. That is, the sphere

center and the total velocity vector (v) are contained in the orbital plane, as shown schematically in Fig. 5. Since the radial line between the sphere center and the vehicle center lies in this plane at all times, the orbital plane is defined alternatively by the sphere center and the normal velocity component (v_ϕ). Consequently, the orbital plane would be rotated by the application of a cross-plane velocity (v_ζ). The orbital plane would then be reorientated in inertial space to contain the sphere center and the vector sum (v_λ) of the normal and cross-plane velocities.

As shown in Fig. 5 by the projection on the unit sphere, the orbit reorientation consists of rotation of the orbital plane about the radial line between sphere center and vehicle center. As the orbital plane is rotated through the angle ζ, the incidence angle, line of nodes and the line of apsides vary from i to λ, Ω to η, and ψ_A to $(\nu - \phi)$, respectively. These variations are defined by the following relations:

$$\zeta = \arctan\left(\frac{v_\zeta}{v_\phi}\right) ; \tag{17}$$

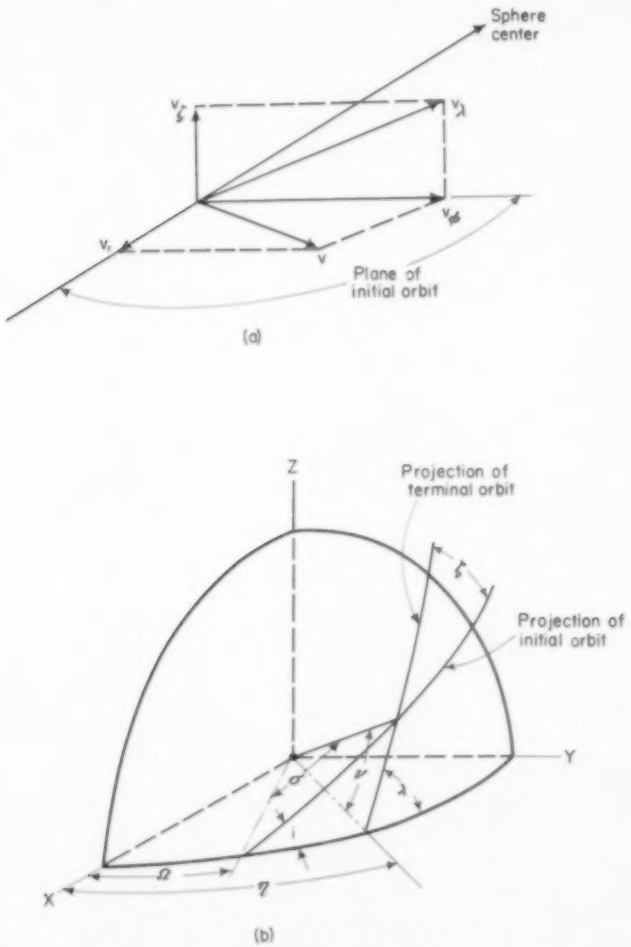


Fig. 5. Orientation of orbital planes:
(a) velocity vectors defining orbital planes; (b) definition of orbital planes in space.

$$\lambda = \arccos\left[\cos i \cos\zeta - \sin i \sin\zeta \cos\sigma\right]; \quad (18)$$

$$\eta = \Omega + \operatorname{arc\,cot}\left[\cos i \cot\sigma + \sin i \cot\zeta \operatorname{cosec}\sigma\right]; \quad (19)$$

$$\nu = \operatorname{arc\,cot}\left[\cos\zeta \cot\sigma + \cot i \sin\zeta \operatorname{cosec}\sigma\right]. \quad (20)$$

The application of a cross-plane velocity as the orbital vehicle passes through the equatorial plane will vary only the incidence angle; the nodal line will not be rotated.

As a result of the vector sum of the normal and cross-plane velocities, the magnitude of the normal velocity in the reorientated orbital plane (B) will differ from that in the original plane (A) in accordance with the following relation:

$$v_{\phi_B} = v_\lambda = [v^2_{\phi_A} + v^2_\zeta]^{1/2}. \quad (21)$$

Also,

$$C_B = C_A\left[1 + \left(\frac{v_\zeta}{v_\phi}\right)^2\right]^{-1/2} \quad (22)$$

and

$$R_B = [v^2_{r_A} + (v_{\phi_B} - C_B)^2]^{1/2}. \quad (23)$$

Since the normal velocity for conservative orbits will usually be much greater than the necessary or practically attainable cross-plane velocity for orbital transfer or reshaping, this velocity magnitude change and the consequent changes of the velocity parameters will probably be negligible in most transfer problems.

In general, the line of apsides will be rotated by reorienting the orbital plane and also by reshaping the orbit in a plane. Consequently, the integral for the difference between the apsidal lines of two orbits will be a function of both the cross-plane velocity and the in-plane velocity change. Although this coupling may complicate the solution of the complete transfer problem, certain restrictions may be imposed upon these transfer velocities to simplify solution.

INTEGRAL FORMULATION OF THE TRANSFER TRAJECTORY

The transfer trajectory must satisfy the following six integral equations for the difference between variations for the mean anomaly ϕ and the parameters i, Ω, ψ, C and R:

$$i_{A_f} - i_{B_f} = (i_{A_0} - i_{B_0}) + \int_{t_0}^{t_f} (\dot\lambda_A - \dot\lambda_B)\,dt = 0 \quad (24)$$

$$\Omega_{A_f} - \Omega_{B_f} = (\Omega_{A_0} - \Omega_{B_0}) + \int_{t_0}^{t_f} (\delta\dot\Omega_{CP} + \delta\dot\Omega_J)\,dt = 0 \quad (25)$$

where

$$\delta\dot\Omega_{CP} = \dot\eta_A - \dot\eta_B$$

$$\delta\dot\Omega_J = \dot\Omega_{A(J)} - \dot\Omega_{B(J)}$$

and

$$\psi_{A_f} - \psi_{B_f} = (\psi_{A_0} - \psi_{B_0}) + \int_{t_0}^{t_f} (\delta\dot\psi_{CP} + \delta\dot\psi_{IP} + \delta\dot\psi_J)\,dt = 0 \quad (26)$$

where

$$\delta\dot\psi_{CP} = (\dot\nu_A - \dot\phi_A) - (\dot\nu_B - \dot\phi_B),$$

$$\delta\dot\psi_{IP} = \dot\psi_{A(RC)} - \dot\psi_{B(RC)}$$

and

$$\delta\dot\psi_J = \dot\psi_{A(J)} - \dot\psi_{B(J)}$$

$$C_{A_f} - C_{B_f} = (C_{A_0} - C_{B_0}) + \int_{t_0}^{t_f} (\dot C_A - \dot C_B)\,dt = 0 \quad (27)$$

$$R_{A_f} - R_{B_f} = (R_{A_0} - R_{B_0}) + \int_{t_0}^{t_f} (\dot R_A - \dot R_B)\,dt = 0 \quad (28)$$

$$\phi_{A_f} - \phi_{B_f} = (\phi_{A_0} - \phi_{B_0}) + \int_{t_0}^{t_f} (\dot\phi_A - \dot\phi_B)\,dt = \pm 2\pi j^* \quad (29)$$

Because the velocity vectors of the satellite and transfer vehicle must be matched at the terminal point for acceptable orbital transfer (or rendezvous), the following terminal conditions should be fulfilled by the transfer trajectory:

$$\dot\lambda_{A_f} = \dot\lambda_{B_f}$$

$$\dot\Omega_{A_f} = \dot\Omega_{B_f}$$

$$\dot\psi_{A_f} = \dot\psi_{B_f}$$

$$\dot C_{A_f} = \dot C_{B_f}$$

$$\dot R_{A_f} = \dot R_{B_f}$$

$$\dot\phi_{A_f} = \dot\phi_{B_f}.$$

*j = any one real integer (0, 1, 2, ...).

Variations of the target satellite parameters as a result of satellite maneuver are provided in the integrands. If the satellite remains in its original reference orbit without maneuver, then

$$\lambda_B = \dot{\eta}_B = \dot{\nu}_B = \dot{\psi}_{B(RC)} = \dot{C}_B = \dot{R}_B = 0$$

and the orbital velocity $\dot{\phi}_B$ is determined by the ballistic state only. Furthermore, correction or reshaping of a satellite orbit without transfer will be formulated with the first five integral equations (equations 24–28), with the simplifying conditions that

$$\lambda_B = \dot{\eta}_B = \dot{\nu}_B = \dot{\psi}_{B(RC)} = \dot{C}_B = \dot{R}_B = 0$$

and

$$\dot{\Omega}_{B(J)} = \dot{\psi}_{B(J)} = 0\,.$$

Alternative forms of the integrand functions for the integral equations may be derived from the equations of the previous sections. For example, integrands for $\dot{\Omega}_J$, $\dot{\psi}_J$, $\dot{C}$ and $\dot{R}$ may be functions of (E,h,m), (C,R), (C,n) or (v_ϕ, v_r). If the eccentricity of orbit is small enough to justify approximation of $\dot{\phi}$ by $\dot{M}$, then the phase relation is reducible to a function of C, $\dot{C}$, R, $\dot{R}$ and t.

Study of the set of integral equations reveals that transfer analysis is complicated by the cross-coupling effects in the nodal equation (equation 25) and the apsidal equation (equation 26). That is, the planar orbit parameters and the in-plane velocity change determine the integrand terms: $\delta\dot{\Omega}_J$, $\delta\dot{\psi}_{IP}$, and $\delta\dot{\psi}_J$. If these terms could be neglected, then the complete transfer problem could be studied by analysis of each of two sets of three integral equations: the incidence, nodal and apsidal line reorientations determined by cross-plane velocity, and the planar orbit parameters and vehicle phase in orbit determined by in-plane velocity change. As an alternative, the apsidal variation due to in-plane velocity change may be pre-assigned, thereby enabling the desired separation of the two sets of three integral equations.

Brief study of the oblateness effects (presented by equations 1 and 2) shows that the individual variations ($\dot{\Omega}_{A(J)}$, $\dot{\Omega}_{B(J)}$, $\dot{\psi}_{A(J)}$, $\dot{\psi}_{B(J)}$) are quite small. For example, the rate of apsidal line rotation ($\dot{\psi}_J$) of an equatorial orbit with nominal circular radius of 5000 statute miles (altitude about 1000 statute miles) will be about 0·001 the mean orbital rate of rotation. Consequently, the oblateness effects could be neglected if the time interval of integration (or problem solution) were short. The permissible time interval for most transfer missions is estimated as 2 hr or less. The permissible time interval for most orbit correction applications will be considerably smaller than for transfer missions. As a result, the oblateness effects will be negligible in such problems, although these effects must be considered in the prediction of ballistic orbits and the start time for transfer.

The apsidal variation due to in-plane velocity change ($\delta\dot{\psi}_{IP}$) may be eliminated by imposing the condition that the time integration of apsidal line changes, over the time interval of solution, must be zero. Although this condition restricts the classes of acceptable solutions severely, a vast field of suitable trajectories are thus attainable with relatively simple analysis.

Since the planar orbit is periodic, the integral equation for orbital phase (equation 29) can be satisfied for any arbitrarily selected value of j. Obviously, a minimum time interval for solution $(t_f - t_0)$ indicates the use of $j = 0$ or 1. However, other design considerations might indicate the use of larger values in certain special cases.

In addition to the orbital transfer and correction problems, a communication satellite or space station must be maintained (or posted) continuously in a given reference orbit. In this case, continuous or discrete correction of the orbit must be provided by a thrust program to eliminate the oblateness effects. The effective time interval for solution would be defined by the period of thrust program recycling. Trajectory analysis for this orbit posting would not be subject to the conditions previously noted for orbital transfer and correction.

THE ACCELERATION HODOGRAPH

The thrust program for a transfer vehicle must be related to the nonballistic normal and radial accelerations (or differential accelerations) required for a transfer trajectory. The hodograph equations (equations 13 and 14) enable the

direct derivation of the relations between the differential accelerations (δa_ϕ, δa_r) and the velocity parameter variations (C,R). As presented previously,

$$v_r = R \sin \phi \tag{13}$$

and

$$v_\phi = C + R \cos \phi \,. \tag{14}$$

Consequently,

$$\dot{v}_r = \dot{R} \sin \phi + R\dot{\phi} \cos \phi \tag{30}$$

and

$$\dot{v}_\phi = \dot{C} + \dot{R} \cos \phi - R\dot{\phi} \sin \phi \,. \tag{31}$$

For a ballistic orbit, $\dot{C} = \dot{R} = 0$, so that

$$\dot{v}_{r(\text{bal})} = R\dot{\phi} \cos \phi \,, \tag{32}$$

$$\dot{v}_{\phi(\text{bal})} = -R\dot{\phi} \sin \phi \tag{33}$$

and

$$\dot{v}_{(\text{bal})} = R\dot{\phi} \,. \tag{34}$$

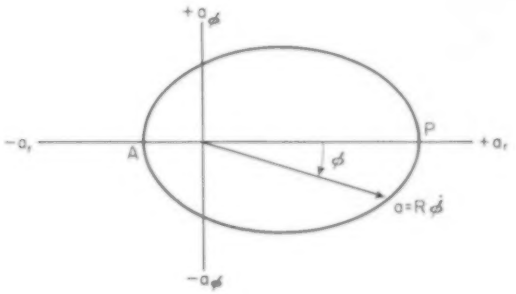
Since the orbital velocity $\dot{\phi}$ varies as an elliptic function in the vehicle orbit, the acceleration hodograph for a ballistic orbit will appear as shown schematically in Fig. 6. The acceleration hodograph of a circular orbit will also be a circle. The differential accelerations are related to the variation of velocity parameters by the equations

$$\dot{v}_{r(\text{pow})} = \delta a_r = \dot{R} \sin \phi \,, \tag{35}$$

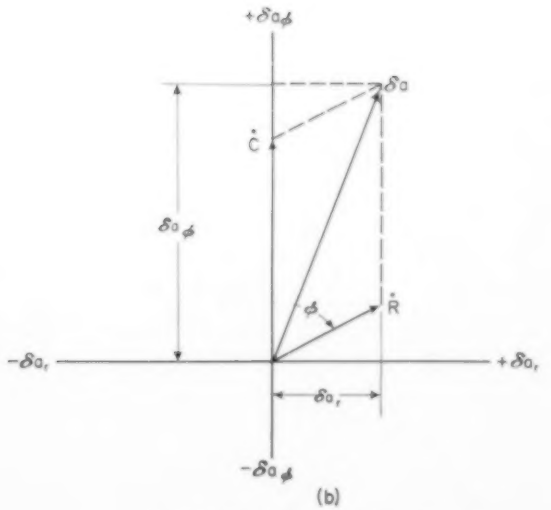
$$\dot{v}_{\phi(\text{pow})} = \delta a_r = \dot{C} + \dot{R} \cos \phi \tag{36}$$

and

$$\delta a_r^2 + (\delta a_\phi - \dot{C})^2 = \dot{R}^2 \,. \tag{37}$$



(a)



(b)

Fig. 6. Acceleration hodograph:
(a) acceleration hodograph for an orbit; (b) differential acceleration for a transfer trajectory

The construction of the differential acceleration hodograph is shown schematically in Fig. 6. The differential acceleration hodograph is invaluable in relating the design requirements for a vehicle transfer propulsion system to a desired transfer trajectory. Note that the algebraic forms of the velocity hodograph orbits A and B at four points each, rather than two points each as shown. However, the solution for the transfer path shown in the figure is of greatest interest because the solution time is the shortest of the maximum number of four possible solutions for a given intermediate orbit and its orientation.

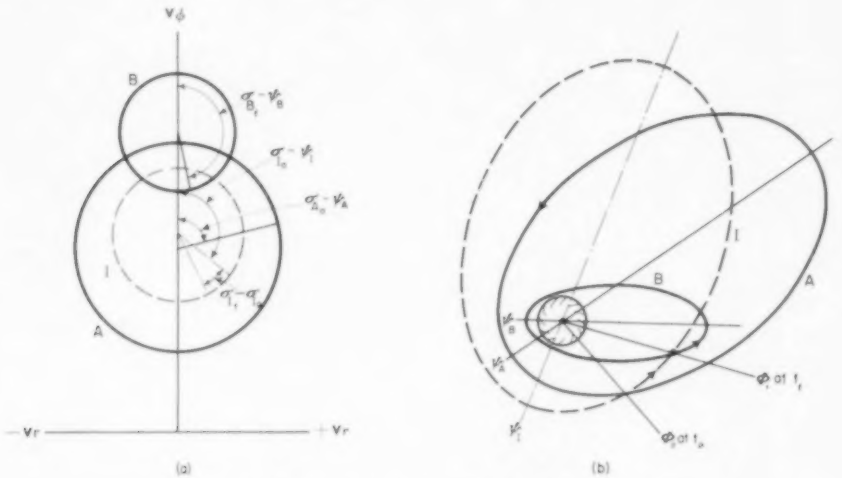


Fig. 7. Two-impulse transfer in orbital plane:
(a) orbital hodographs; (b) orbits in space.

equation (equation 15) and the differential acceleration hodograph equation (equation 37) are identical.

IMPULSIVE TRANSFER IN AN ORBITAL PLANE

Planar orbit transfer by one or two step velocity changes, produced by acceleration impulses, has been continually analysed and discussed for many decades since Hohmann's work.[5] This general class of problem is simply formulated by use of the polar hodograph.

The transfer between two coplanar, nonintersecting orbits A and B, shown schematically in Fig. 7, is most directly provided by the use of an intermediate orbit I which intersects both orbits A and B. Although the shape and orientation of the intermediate orbit can be arbitrarily selected, all orbits are fixed in inertial space, subject only to vehicle planar transfer changes. The intermediate orbit might intersect both

The transfer points on the polar hodographs are subject to the physical constraint that position in space cannot change with an instantaneous velocity change. Consequently,

$$r_{A_0} = r_{I_0}\,; \quad \Phi_{A_0} = \Phi_{I_0} \tag{38}$$

and

$$r_{I_f} = r_{B_f}\,; \quad \Phi_{I_f} = \Phi_{B_f}\,, \tag{39}$$

or

$$C_A\, v_{\phi A_0} = C_I\, v_{\phi I_0}\,; \psi_A + \int_{t_{\phi A}}^{t_0} \dot{\phi}_A \,\mathrm{d}t = \psi_I + \int_{t_{\phi I}}^{t_0} \dot{\phi}_I \,\mathrm{d}t \tag{38a}$$

and

$$C_I\, v_{\phi I_f} = C_B\, v_{\phi B_f}\,; \quad \psi_I + \int_{t_{\phi I}}^{t_f} \dot{\phi}_I \,\mathrm{d}t = \psi_B + \int_{t_{\phi B}}^{t_f} \dot{\phi}_B \,\mathrm{d}t\,. \tag{39a}$$

Therefore,

$$C_A[C_A + R_A \cos(\sigma_{A_0} - \psi_A)] = C_I[C_I + R_I \cos(\sigma_{I_0} - \psi_I)] \quad (40)$$

and

$$C_I[C_I + R_I \cos(\sigma_{I_f} - \psi_I)] = C_B[C_B + R_B \cos(\sigma_{B_f} - \psi_B)] . \quad (41)$$

It can be shown that

$$\dot{\phi} = \left(\frac{C}{\mu}\right) v_\phi^2 \quad (42)$$

so that

$$\psi_A + \int_{t_{\phi A}}^{t_0} \left(\frac{C_A}{\mu}\right) v_{\phi A}^2 \, dt = \psi_I + \int_{t_{\phi I}}^{t_0} \left(\frac{C_I}{\mu}\right) v_{\phi I}^2 \, dt \quad (43)$$

and

$$\psi_I + \int_{t_{\phi I}}^{t_f} \left(\frac{C_I}{\mu}\right) v_{\phi I}^2 \, dt = \psi_B + \int_{t_{\phi B}}^{t_f} \left(\frac{C_B}{\mu}\right) v_{\phi B}^2 \, dt , \quad (44)$$

or

$$\psi_A + \int_{t_{\phi A}}^{t_0} \frac{C_A}{\mu} [C_A + R_A \cos(\sigma_{A_0} - \psi_A)]^2 \, dt = \psi_I + \int_{t_{\phi I}}^{t_0} \frac{C_I}{\mu} [C_I + R_I \cos(\sigma_{I_0} - \psi_I)]^2 \, dt \quad (43a)$$

and

$$\psi_I + \int_{t_{\phi I}}^{t_f} \frac{C_I}{\mu} [C_I + R_I \cos(\sigma_{I_f} - \psi_I)]^2 \, dt = \psi_B + \int_{t_{\phi B}}^{t_f} \frac{C_B}{\mu} [C_B + R_B \cos(\sigma_{B_f} - \psi_B)]^2 \, dt . \quad (44a)$$

Either set of equations (equations 38 and 39 or 40, 41, 43 and 44) defines the transfer trajectory.

Although these equations are nonlinear, the special case of Hohmann's transfer at the apsidal points reduces to a simple and direct solution. Hohmann transfer consists of an elliptical transfer trajectory tangential to the initial and final orbits at the apsidal points, as shown in Fig. 8. In this case, the major axes of all elliptical orbits must be coincident. Then

$$C_H = C_A \left(\frac{v_1}{v_1'}\right) = C_B \left(\frac{v_4}{v_4'}\right) \quad (45)$$

and

$$R_H = v_1' - C_H = C_H - v_4' , \quad (46)$$

so that

$$C_H = [\tfrac{1}{2}(C_A v_1 + C_B v_4)]^{1/2} \quad (47)$$

and

$$R_H = (C_A v_1 - C_B v_4)[2(C_A v_1 + C_B v_4)]^{1/2} . \quad (48)$$

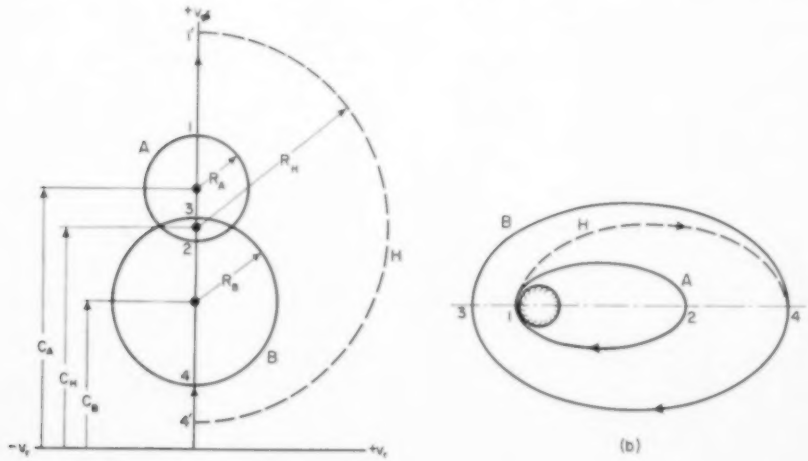


Fig. 8. Hohmann transfer:
(a) orbital hodographs; (b) orbits in space.

Also, the ratio of the required velocity changes to the required velocity at the respective apsidal points is

$$\frac{v_1' - v_1}{v_1} = \frac{C_A}{C_H} - 1 \tag{49}$$

and

$$\frac{v_4 - v_4'}{v_4} = 1 - \frac{C_B}{C_H}. \tag{50}$$

Since the transfer time Δt is one-half of the transfer orbit period,

$$\Delta t = \frac{\pi}{n_H} = \mu\pi\left[\frac{1}{2}\left(\frac{1}{C_A v_1} + \frac{1}{C_B v_4}\right)\right]^{3/2} \tag{51}$$

and the orbital phase (referred to the terminal transfer point) of a target satellite at the start time of the transfer must be

$$\Delta\phi = \pi\left[\frac{1}{2}(C_B^2 - R_B^2)\left(\frac{1}{C_A v_1} + \frac{1}{C_A v_4}\right)\right]^{3/2} \tag{52}$$

or

$$\Delta\phi = \pi\left[\frac{C_B^2 - R_B^2}{C_H^2 - R_H^2}\right]^{3/2}. \tag{52a}$$

Impulsive transfer in the orbital plane, as presented in this section, has implied that the entire apsidal line difference is eliminated by in-plane velocity change. Any concurrent cross-plane velocity required for orbital plane re-orientation would then be restricted to a zero total change in apsidal line position. Consequently, it is apparent that optimalization in the orbital plane alone will not necessarily provide optimalization of the complete transfer solution.

CONCLUSIONS

The missions of man-made satellites will require orbital transfer, correction (or reshaping) of orbit or posting in orbit. The transfer trajectory in three-dimensional space is formulated by a set of six integral equations with six parameters: i, Ω, ψ, ϕ, C, R. The two velocity parameters C and R replace the conic parameters a and e for description of the planar orbit. The six integral equations may be separated into two sets of three equations each for analysis, whenever

(1) the cross-plane velocity is much smaller than the in-plane velocity;

(2) the oblateness effects on nodal and apsidal line rotation are negligible due to a relatively short time interval of integration for solution; and

(3) the apsidal line rotation due to in-plane velocity change over the time interval of integration is pre-assigned.

This formulation enables the optimalization for such items as minimum time and fuel consumption in the three-dimensional transfer problem.

The application of the complete transfer analysis to the engineering design of a transfer guidance system will be different for the mid-course and terminal phases. Engineering mechanization for mid-course guidance will probably be based on ground observations and the consequent derivation of the six parameters explicitly. However, terminal guidance will undoubtedly use observational data, derived at the transfer vehicle, of the relative bearing and range to the satellite and possibly of the vehicle local horizon. Consequently, the six parameters will be available in implicit form, in this case.

Although the formulation presented in this paper is intended for orbital transfer or correction trajectories, the basic integral presentation should be effectively applicable to the orbital intercept problem. In this case, arrival in the proximity of the target satellite constitutes solution, rather than coincidence of spacial and velocity co-ordinates. Also, such arrival at the terminal volume of space may be accomplished by a trajectory that need not approach the orbit of the target satellite tangentially. The required terminal conditions for successful intercept will be determined largely by the design performance of the interception system and the target end-game techniques of orbital guidance.

Nomenclature

a = semi-major axis of the ellipse (h)
A = apogee of the orbit
C = velocity parameter of the orbit (ht^{-1})
e = eccentricity of the ellipse

E = total potential and kinetic energy of the orbital body (mh^2t^{-2})
$\bar{E}$ = total orbital energy per unit mass of the orbital body (h^2t^{-2})
h = total angular momentum of the orbital body (mh^2t^{-1})
i = incidence angle between the orbital plane and the equatorial plane of the celestial body
j = any one real integer (0, 1, 2, . . .)
J = first coupling constant of the spherical harmonic expansion for the gravitational potential of the celestial body, described as an oblate spheroid
m = mass of the orbital body (m)
M = mean anomaly of the body in orbit
n = mean angular velocity of the body in orbit (t^{-1})
P = perigee of the orbit
r = radial distance between the gravitational centers of theor bital and celestial bodies (h)
R = velocity parameter of the orbit (ht^{-1})
R_e = equatorial radius of the Earth (h)
t = time (t)
T = time of perigee passage in orbit (t)
v = linear velocity (ht^{-1})
δa = differential acceleration due to thrust (ht^{-2})
Δt = interval of time (t)
β = angle between the hodograph v_ϕ-axis and the hodograph total velocity
γ = angle between the hodograph v_ϕ-axis and the hodograph radius (R) to the orbital body on the hodograph
ζ = angle of orbital plane rotation
η = reference angle of the nodal line for the rotated orbital plane
θ = angle between the total velocity vector and the local horizon in the orbital plane
κ = coefficient of gravitational attraction between the orbital and celestial bodies = $\mu m(mh^3t^{-2})$
λ = incidence angle between the rotated orbital plane and the equatorial plane of the celestial body
μ = terrestrial gravitational constant (h^3t^{-2})
ν = reference angle between the nodal line for the rotated orbital plane and the line between centers of the orbital and celestial bodies
π = numerical constant (3·14159 . . .)
σ = angle between the nodal line of the orbital plane and the radial line between centers of the orbital and celestial bodies
ϕ = true anomaly of the body in orbit
Φ = the radial line between centers of the orbital and celestial bodies
ψ = angle between the nodal line of the orbital plane and the apsidal line of the orbit at perigee
Ψ = the apsidal line of the orbit at perigee
Ω = reference angle of the nodal line for the orbital plane

Subscripts

A = initial orbit
B = final orbit
H = Hohmann transfer ellipse
I = intermediate transfer ellipse
J = oblateness effects
f = final state
0 = initial state
CP,ζ = cross-plane
IP = in-plane
RC = due to in-plane velocity change
r = radial
bal = ballistic
pow = powered
ϕ = tangential

REFERENCES

1. D. Brouwer, *Astronom. J.* No. 1156, p. 223 (1946).
2. A. A. Orlov, *Pochti krugovye periodicheskie dvizheniya material 'noi tochki pod deistviem nyutonovskogo prityazheniya sferoida* (Almost Circular Periodic Motions of a Particle of Matter under the Gravitational Attraction of a Spheroid), Translation No. 519 by A. Pingell, Naval Res. Lab. (1955).
3. L. Blitzer, M. Weisfeld and A. D. Wheelon, *J. Appl. Phys.* **27**, 1141 (1956).
4. L. Blitzer, *J. Appl. Phys.* **28**, 1362 (1957).
5. W. Hohmann, *Die Erreichbarkeit der Himmelskörper*, Oldenbourg, Munich (1925).

TRAJECTORIES FOR CONTINUOUSLY APPLIED PROPULSIVE THRUST

J. VICTOR HUGHES and GEORGE N. NOMICOS
Scientific Research Staff, Republic Aviation Corporation

Abstract—The amount of propellant consumed by a low-thrust interplanetary vehicle in accomplishing a specified mission is investigated. It is shown that *the "characteristic velocity" for transfer from one circular orbit to another, via a spiral path under continuously applied thrust, is just equal to the difference between the velocities in the two circular orbits.*

The result is proved rigorously for transfer via an equi-angular spiral, and the result is also shown to be valid for any spiral path, so long as the velocity of the vehicle at any point of the path can be equated to the velocity in a circular orbit passing through the point.

A second approximation is introduced, in which the angle of the spiral is allowed to vary. With this approximation, it is shown that the time taken to perform a mission is equal to the characteristic velocity divided by the thrust-to-mass ratio. In conclusion, some numerical results for a typical mission are evaluated.

1. INTRODUCTION

For interplanetary missions it is common to assume the use of an engine giving very high specific impulse, because this conserves propellant. Examples of such engines are the plasma engine,[1, 2] the ion engine[3] and the photon engine.[4] Such engines all give low thrust, and consequently have to operate for long periods to give an appreciable change of velocity; it is common to assume the engine will operate for all, or nearly all, of the time taken for the mission.

Such a low-thrust engine cannot lift the vehicle off the Earth, so it is usually assumed that the mission starts from an orbit round the earth. Similarly the engine thrust is insufficient to prevent catastrophe if the vehicle approaches a planet too closely, so it is usually assumed that the mission terminates in an orbit round a planet. Thus a typical mission for a vehicle with low-thrust engine might be

(a) starting from an orbit round the Earth, operate the engine to cause the vehicle to spiral outward until eventually it succeeds in escaping the Earth's attraction;
(b) operate the engine to cause transfer from the Earth's heliocentric orbit to the heliocentric orbit of some specified planet;
(c) operate the engine to cause the vehicle to spiral inwards towards (and round) the planet, until eventually it stabilizes in an orbit round the planet.

Missions of this sort, from Earth to Mars, have been examined in a general way,[5, 6] but the characteristic velocity necessary for such a mission, when using a low-thrust engine, has not been adequately studied. Since characteristic velocity is a direct measure of propellant consumption, this is a very important matter. It is shown in this paper that the characteristic velocity for a low-thrust engine on such a mission is between two and three times the characteristic velocity required by a very high-thrust engine. In fairness to the low-thrust engine, it should be pointed out that the long burning time gives a possibility of trajectory correction which is much more flexible than for the engine giving a short period of high thrust.

In calculating the characteristic velocity, the motion of a vehicle on an equi-angular spiral under the action of a single attracting center is first studied, and it is shown that the characteristic velocity to transfer from one radius to another is exactly equal to the difference between the velocity in a circular orbit through the initial point and that for a circular orbit through the final point. For the mission described above, from an orbit round the Earth to an orbit round

Mars, the three sections are treated separately, applying the theory to each as though the other attracting centers were absent. This is a reasonable approximation which has been adopted in many calculations.

One result of the equi-angular spiral is that the thrust must decrease as the distance from the attracting center (Earth, Sun or planet) increases. In a second approximation, the thrust is kept constant while the angle of the spiral varies, and under these conditions the time of transit is shown to be equal to the characteristic velocity divided by the thrust-to-mass ratio. As a demonstration of these results, some numerical results are calculated for a typical mission.

2. DEFINITION OF "CHARACTERISTIC VELOCITY"

Characteristic velocity was originally defined for rockets whose thrust was large compared to the other forces acting on the vehicle. Under these circumstances, the characteristic velocity was defined as the change in vehicle velocity during the burning time. For a mission requiring two or more short intervals of burning, the characteristic velocity for the entire mission was obtained by adding the velocity changes which occurred during the several burning times.

Newton's law applied to the rocket, integrated over the time of burning, is given in equation (1)*

$$\Delta v = \int \frac{T}{m}\,\mathrm{d}t + \int \frac{Y}{m}\,\mathrm{d}t \tag{1}$$

For the condition of thrust greatly exceeding external forces, for which the characteristic velocity V is defined as equal to Δv, the last term of this equation can be omitted, which gives equation (2):

$$V = \int \frac{T}{m}\,\mathrm{d}t = \int p\,\mathrm{d}t \tag{2}$$

The thrust is related to the rate of emergence of propellant, and propellant velocity, by

$$T = -U\frac{\mathrm{d}m}{\mathrm{d}t} \tag{3}$$

* For nomenclature see end of paper.

Combining this with equation (2) yields the well-known equation

$$V = U \ln \frac{m_0}{m_0 - m_p} \tag{4}$$

The importance of characteristic velocity is as a convenient measure of propellant consumption, as expressed by equation (4). For the case of low-thrust engines this relation must be retained, which means that equation (2) is used as the definition of characteristic velocity for the case of a low-thrust as well as for a high-thrust engine. For the low-thrust case, in equation (1) the external forces Y may well be much greater than the thrust T, so that the velocity change Δv, as calculated from equation (1), will in this case have no relation to the characteristic velocity V. For the low-thrust case, equation (2) is the most convenient form for calculating characteristic velocity.

3. MOTION IN AN EQUI-ANGULAR SPIRAL

(A) The General Equations

A vehicle spiralling slowly out from an orbit round an attracting center (Sun, Earth or other planet) will follow a path which for analysis puposes can be treated as an equi-angular spiral. Such a spiral, in which the flight path always makes the same angle with the radius vector, is sketched in Fig. 1, and an enlarged portion is shown in Fig. 2. From this figure, and the definition of k:

$$k = \tan\beta = \frac{1}{r}\frac{\mathrm{d}r}{\mathrm{d}\theta} \tag{5}$$

Integration yields the equation of the spiral:

$$r = r_0 \mathrm{e}^{k\theta} \tag{6}$$

If the thrust of the rocket vehicle, acting alone, could give to the vehicle an acceleration p (not necessarily a constant), and if this thrust is directed along the flight path, then the equations of motion are

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} - r\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 = p\sin\beta - \frac{\mu}{r^2}$$

$$r\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + 2\frac{\mathrm{d}r}{\mathrm{d}t}\frac{\mathrm{d}\theta}{\mathrm{d}t} = p\cos\beta$$

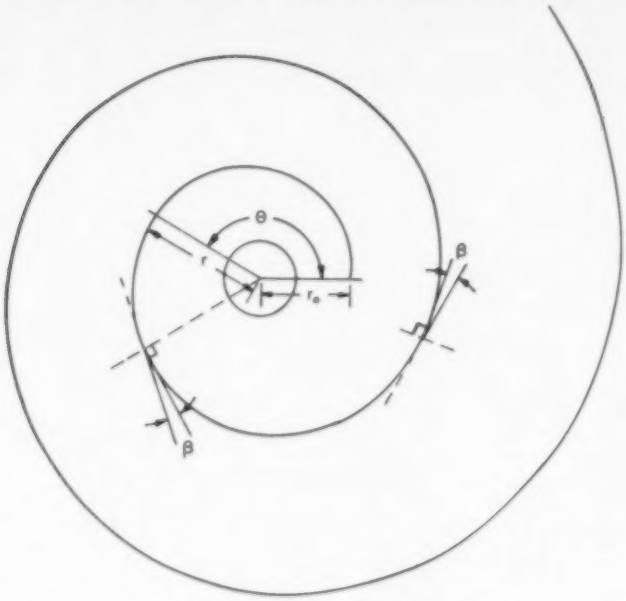


Fig. 1. The equi-angular spiral.

It follows, from the definition of ω, that any time derivative can be replaced as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t} \equiv \omega \frac{\mathrm{d}}{\mathrm{d}\theta}$$

Using this and equation (5) allows the equations of motion to be written

$$kr\omega \frac{\mathrm{d}\omega}{\mathrm{d}\theta} - r\omega^2(1 - k^2) = p \sin \beta - \frac{\mu}{r^2} \qquad (7)$$

$$r\omega \frac{\mathrm{d}\omega}{\mathrm{d}\theta} + 2kr\omega^2 = p \cos \beta \qquad (8)$$

It is convenient to introduce a new variable X, the centrifugal acceleration, defined by

$$X = r\omega^2 \qquad (9)$$

Direct differentiation, with equation (5), yields:

$$\frac{\mathrm{d}X}{\mathrm{d}\theta} = kX + 2r\omega \frac{\mathrm{d}\omega}{\mathrm{d}\theta}$$

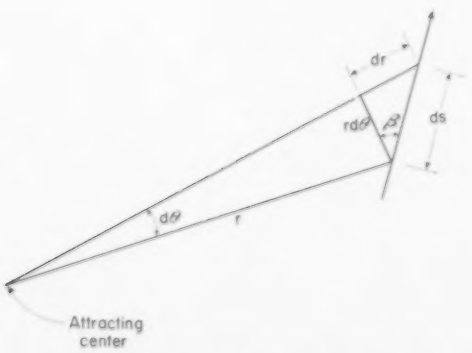


Fig. 2. An enlarged portion of the spiral.

Substituting in equations (7) and (8) gives

$$\frac{1}{2}k\frac{dX}{d\theta} - X\left(1 - \frac{k^2}{2}\right) = p\sin\beta - \frac{\mu}{r^2} \tag{10}$$

$$\frac{1}{2}\frac{dX}{d\theta} + \frac{3}{2}kX = p\cos\beta \tag{11}$$

Subtracting equation (10) from k times equation (11) gives

$$X(1 + k^2) = pk\cos\beta - p\sin\beta + \mu/r^2 = \mu/r^2$$

because the terms in β vanish from the definition of k. This equation can be written

$$X = \frac{\mu}{r^2(1 + k^2)} \tag{12}$$

If equation (6) is substituted in equation (12), differentiation with respect to θ is possible, and the left-hand side of equation (11) can be evaluated. This gives:

$$p\cos\beta = \frac{1}{2}\frac{k}{1 + k^2}\frac{\mu}{r_0^2}\exp(-2k\theta) \tag{13}$$

Using equations (5) and (6), this can be written in either of the following forms:

$$p = \frac{1}{2}\frac{k}{\sqrt{(1 + k^2)}}\frac{\mu}{r_0^2}\exp(-2k\theta) \tag{14}$$

$$p = \frac{1}{2}\frac{\mu}{r^2}\sin\beta \tag{15}$$

Equation (15) states that, to follow an equi-angular spiral, the thrust must be programmed to vary inversely as the square of the distance from the attracting center. Another way of viewing this equation is to state that the thrust must be programmed so that at all times it is only one half of the component of gravitational acceleration along the tangent to the flight path. There is thus always a net component of force backwards along the flight path, which slows the vehicle down as it spirals outward.

The velocity at any point in the spiral can be found by noting from Fig. 2:

$$v = \frac{ds}{dt} = \frac{r\,d\theta/\cos\beta}{dt} = \frac{r\omega}{\cos\beta}$$

Using equations (9) and (12) to evaluate $r\omega$ yields:

$$v = \sqrt{\left(\frac{rX}{\cos^2\beta}\right)} = \sqrt{\frac{\mu}{r}} = v_c \tag{16}$$

This gives the interesting result that the velocity at any point in the spiral is exactly equal to the velocity in a circular orbit passing through that point.

The time for any transfer can readily be found. Combining equations (6), (9) and (12) yields

$$\frac{d\theta}{dt} = \left[\frac{\mu}{r_0^3(1 + k^2)}\right]^{1/2}\exp(-3k\theta/2) \tag{17}$$

Integration, making use of equation (6), gives

$$t = \frac{2}{3\sin\beta}\frac{1}{\mu^{1/2}}(r_f^{3/2} - r_0^{3/2})$$

or

$$t = \frac{2}{3\sin\beta}\left(\frac{r_f}{v_{cf}} - \frac{r_0}{v_{c0}}\right) \tag{18}$$

The angle turned through by the radius vector during this transfer is obtained directly from equation (6):

$$\theta = \frac{1}{k}\ln(r_f/r_0) \tag{19}$$

Most of these results have been published independently.[7,8] An alternative treatment leading to the same results, which uses resolution of accelerations along and perpendicular to the path, is given in the appendix.

(B) The Characteristic Velocity

For transfer between any two prescribed radii, by means of an equi-angular spiral, the characteristic velocity can be obtained as follows:
From equation (4),

$$V = \int p\,dt = \int \frac{p\,d\theta}{d\theta/dt}$$

Substitution from equations (14) and (17) gives

$$V = \int\frac{k}{2}\sqrt{\frac{\mu}{r_0}}\exp(-k\theta/2)\,d\theta$$

$$= \sqrt{\frac{\mu}{r_0}}[1 - \exp(-k\theta/2)]$$

$$= \sqrt{\frac{\mu}{r_0}} - \sqrt{\frac{\mu}{r_f}}$$

Thus

$$V = v_{c0} - v_{cf} \tag{20}$$

Thus the characteristic velocity for transfer between two circular orbits, using continuous firing of a low-thrust engine, is just equal to the difference between the speeds in the two orbits. In particular, for a vehicle leaving a circular orbit round a planet and spiralling out to infinity, the characteristic velocity required is equal to the original vehicle velocity in the circular orbit.

4. ESCAPE FROM CIRCULAR ORBIT: ALTERNATIVE TREATMENT

The same result as in the previous section, that the characteristic velocity for slow change from one circular orbit to another is equal to the difference of the velocities in the orbits, can be proved without reference to the specific transfer path, as follows.

The work done by the thrust T acting along a length of path ds is converted to an increase of energy of the vehicle. Dividing by vehicle mass, this gives

$$p\,ds = dE \tag{21}$$

From equation (4),

$$V = \int p\,dt = \int \frac{p\,dt}{ds}\,ds = \int \frac{p\,ds}{v} = \int \frac{dE}{v} \tag{22}$$

Equation (22) is a general relation, provided only that the thrust is tangential to the path, so that equation (21) is valid, but the evaluation of the integral depends on the conditions. For a sudden impulse, the change in energy is all kinetic. For the spiral, the change is partly kinetic and partly potential. For a very small angle spiral, we may reasonably assume the total energy at any point to be the same as for a circular orbit passing through the point. For the equi-angular spiral this was shown earlier to be rigorously true, since the velocities, and hence the kinetic energies, in the spiral and the circular orbits were exactly equal, and of course the potential energy is fixed for a given point. For a circular orbit of radius r_c

$$E = -\mu/2r_c \tag{23}$$

$$v = \sqrt{(\mu/r_c)} \tag{24}$$

Differentiating equation (23) and substituting in equation (22) gives

$$V = \int \frac{\mu\,dr_c/2r_c^2}{\sqrt{(\mu/r_c)}}$$

$$= \sqrt{\frac{\mu}{r_0}} - \sqrt{\frac{\mu}{r_f}}$$

or

$$V = v_{c0} - v_{cf}$$

This agrees with the earlier equation (20). The only restriction is that the total energy, at any point in the path, must equal that for a circular orbit through the point. This will be true if the spiral angle is very small (since then the spiral will closely approximate a circle); and it may be true in some special cases, such as the case of an equi-angular spiral, even if the spiral angle is not small.

5. A SECOND APPROXIMATION

In the above, the theory was developed for motion in a spiral, with the angle between the path and the radius vector from the attracting center constant. The consequence of this was that the thrust had to decrease as the distance from the attracting center increased, as expressed in equation (15). But a more desirable program would be to maintain the thrust constant at its maximum value, allowing the angle of the spiral to increase as the distance from the attracting center increased. This means that equation (15) would be considered as giving the angle β as a function of the distance r. This is not strictly in accordance with the early theory, since the theory treated β as a constant, but if the rate of change of β is sufficiently small the approximation should be acceptable. The criterion by which the smallness of the rate of change of β is to be judged will be the subject of a later paper.

Using the original equations for the equi-angular spiral, the time to transfer from one radius to another was found in equation (18). But if the second radius is large compared to the first, it will be found that equation (18) gives an excessively long time of transit. This is because, at the larger values of radius, the thrust has decreased very considerably. If the alternative program is assumed, in which the angle of the

spiral varies while the thrust is maintained constant, a much shorter transfer time is found. In equation (2), if the thrust p, is maintained constant, the integral becomes merely pt, and the simple result is obtained:

$$t = V/p \tag{25}$$

Due to the approximation involved, this time, using V as found from equation (20), is an underestimate, but in practical cases it is unlikely to be in error by more than a few per cent.

6. NUMERICAL VALUES

In Ref. (1) a space vehicle powered by a plasma engine was designed to have a thrust-to-weight ratio of $0{\cdot}75 \times 10^{-4}$. Ref. (3), describing a space vehicle with ion engine, gave a thrust-to-weight ratio of the same order of magnitude. From these examples, a typical space craft with low-thrust engine might be expected to have an acceleration of about 10^{-4} of the gravitational acceleration at the Earth's surface, or 0·098 cm/sec^2.

A typical mission will be considered in which the vehicle starts from a circular orbit round the Earth, at an altitude of 1000 km, and terminates in an orbit round Mars at an altitude of 500 km. The constants, and the velocities involved, are shown in Table 1. Various aspects of this mission are discussed in more detail below.

Table 1. Constants for Orbits round Earth, Sun and Mars

Attracting center	Constant μ, (km^3/sec^2)	Radius of orbit (km)	Velocity in orbit (km/sec)	Velocity change (km/sec)
Earth	$3{\cdot}99 \times 10^5$	$r_0 = 7400$	7·35	
		$r_f = \infty$	0	7·35
Sun	$1{\cdot}33 \times 10^{11}$	$r_0 = 1{\cdot}50 \times 10^8$	29·80	
		$r_f = 2{\cdot}28 \times 10^8$	24·14	5·66
Mars	$4{\cdot}30 \times 10^4$	$r_f = \infty$	0	
		$r_0 = 3900$	3·32	3·32
Total				16·33

(A) Characteristic Velocity

It will be seen from Table 1 that the characteristic velocity for this mission, using this low-thrust engine, is 16·33 km/sec. On the other hand, a vehicle using a very high thrust engine could, in theory at least, accomplish the transfer using just two impulses; one of these impulses, of 3·45 km/sec, given while in orbit round the Earth, would put the vehicle in a hyperbolic trajectory, escaping from the Earth's field with a velocity sufficient to establish the vehicle in a Hohman transfer ellipse; if conditions were chosen precisely correctly, at aphelion of this ellipse the vehicle would enter a hyperbolic orbit round Mars, and an impulse of 2·09 km/sec, at the point of closest approach, would establish the vehicle in orbit round Mars. Thus the total characteristic velocity for transfer, using the high-thrust engine, is only 5·54 km/sec, or just over one-third of that needed by the low-thrust engine.

It will be seen, from the above description, that exceedingly great accuracy would be needed to carry out the maneuver outlined for the high-thrust transfer. In practice, corrective maneuvers would certainly be required, which could be quite expensive in propellant (i.e. characteristic velocity). By contrast, the low-thrust engine, operating most or all of the time of transfer, can execute corrections as necessary during flight in a very economic manner. Thus the possibility of continuous trajectory correction may make up for the apparent advantage of the high-thrust engine in requiring less characteristic velocity.

(B) Escape from Geocentric Orbit

In an orbit round the Earth at low altitude, where μ/r^2 can be taken as equal to g_0, the numerical values give β about 2×10^{-4} rad, or about 0·01°. If the thrust is kept constant, $\sin\beta$ must vary as r^2, which leads to the values listed in Table 2. It will be seen that the angle has become 90°, which means that the vehicle is travelling radially outward, at approximately the distance of the Moon. At this point the vehicle can be said to have escaped. The gravitational attraction at this point is still twice the

thrust, but the momentum appropriate to a circular orbit through this point is sufficient to overcome the difference.

On the question of entry into the spiral, it will be seen that the angle of the spiral near the Earth

Table 2. Angle of Spiral as Function of Distance from Earth

Approx. height above Earth (miles)	Approx. radius of orbit (miles)	Approx. angle of spiral (degrees)
100	4,000	0·01
8,000	12,000	0·1
36,000	40,000	1·0
120,000	120,000	10·0
280,000	280,000	90

is only 2×10^{-4} rad. Putting into equation (6) the values $r_0 = 4000$ miles, $\theta = 2\pi$ rad, $k = 2 \times 10^{-4}$, yields (on expanding the exponential)

$$(r_1 - r_0) = 5 \text{ miles}$$

so that one complete revolution round the Earth gives an increase of radius of only 5 miles. Correspondingly, for the simple theory to apply as given, it would be necessary for the take-off orbit to be a circle within about ± 2 miles. This makes it clear that the problem of take-off by a low-thrust engine from an elliptic orbit merits further study.

(C) Time of Transit

Assuming the thrust to be maintained constant, and the spiral angle to vary as necessary, equation (25) gives the times for each sector of the journey. The results are shown in Table 3. It will be seen that a total transit time of only 189 days is obtained; even allowing for errors due to the approximations, it seems that a total journey time of about 200 days should result.

By contrast, using constant angle and maximum thrust when in orbit round the Earth at 1000 km, equation (18) gives about 40 years for spiralling out to the Moon. For the heliocentric spiral, if maximum thrust at the Earth's distance is assumed, and a constant angle, the time to transfer to Mars' distance from the Sun is 103 days, compared to 66 days of Table 3.

Table 3. Times of Transit over Sections of Trip from Geocentric Orbit to Circum-Martian Orbit

Portion of path	V (km/sec)	t (days)
Start at 1000 km above Earth, and spiral out to infinity	7·35	85
Transfer from Earth's heliocentric orbit to Mars' heliocentric orbit	5·66	65·5
Start at infinity and spiral in to an orbit 500 km above Mars	3·32	38·5
Total		189

(D) Summary of Numerical Results

(a) To enter an equi-angular spiral from an orbit round the Earth, a vehicle with a thrust-to-weight ratio of 10^{-4} would need to start from an orbit circular to within ± 2 miles.

(b) The spiral would initially have an angle of about 0·01° to the circular path. As it got progressively farther from the Earth, the angle could increase, as given in Table 2.

(c) In its first revolution round the Earth, the distance from the Earth would be increased by 5 miles. In the first day, the increase in distance from the Earth would be about 40 miles.

(d) Assuming the thrust maintained constant, the times calculated for the various sectors are given in Table 3. Even allowing for errors due to approximations, the total time for the journey would probably be less than 200 days.

(e) The characteristic velocity for a complete mission, by low-thrust engine, from an orbit round the Earth at 1000-km altitude to an orbit round Mars at 500-km altitude is 16·33 km/sec.

7. CONCLUSION

An inter-planetary vehicle with a low-thrust engine, moving in the gravitational field of a single attracting center, can move on an equi-angular spiral under certain conditions, which

are discussed. If the angle of the spiral is kept rigorously constant, the thrust must be programmed to vary as the inverse square of the distance from the attracting center; but in all practical cases, the distance from the attracting center will vary only slowly, and it is probably permissible to keep the thrust constant and allow the angle of the spiral to vary.

It is shown that for motion in such a spiral the characteristic velocity needed to move from one radius to another is just equal to the difference between the speeds in the circular orbits of these radii. It is also shown that the time taken for any transfer, if the thrust is maintained at its maximum value, is given by the characteristic velocity divided by the thrust-to-mass ratio.

Throughout the discussion it has been assumed that the vehicle takes off from a circular or very nearly circular orbit. The question of take-off by a vehicle with a low-thrust engine from an elliptic orbit must be the subject of another paper. Work on the more accurate delineation of the conditions under which variation of the angle of the spiral can be tolerated will also be considered in another paper.

APPENDIX

An Alternative Treatment of the Equi-angular Spiral

In this treatment, acceleration will be resolved along and perpendicular to the flight path. Resolving perpendicular to the path gives

$$\frac{\mu}{r^2}\cos\beta = v^2/R \tag{A1}$$

where R is the radius of curvature. Now

$$R = \mathrm{d}s/\mathrm{d}\psi$$

where $\mathrm{d}s$ is an element of the path and $\mathrm{d}\psi$ the angle through which the tangent to the path has turned while traversing $\mathrm{d}s$. Since the path is at a fixed angle to the radius vector, $\mathrm{d}\theta = \mathrm{d}\psi$. From Fig. 2,

$$\mathrm{d}s = r\mathrm{d}\theta/\cos\beta \tag{A2}$$

Hence

$$R = r/\cos\beta$$

Substituting in equation (A1),

$$v^2/r = \mu/r^2 \tag{A3}$$

Hence

$$v = (\mu/r)^{1/2} \tag{A4}$$

which is the same as equation (16).
Resolving along the flight path

$$p - \frac{\mu}{r^2}\sin\beta = \frac{\mathrm{d}v}{\mathrm{d}t} \tag{A5}$$

Now

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}r}\frac{\mathrm{d}r}{\mathrm{d}t}$$

From Fig. 2,

$$\frac{\mathrm{d}t}{\mathrm{d}r} = \frac{\mathrm{d}s\sin\beta}{\mathrm{d}t} = v\sin\beta$$

Thus

$$\frac{\mathrm{d}v}{\mathrm{d}t} = v\sin\beta\frac{\mathrm{d}v}{\mathrm{d}r} = \frac{1}{2}\sin\beta\frac{\mathrm{d}(v^2)}{\mathrm{d}r}$$

From equation (A4),

$$\frac{\mathrm{d}(v^2)}{\mathrm{d}r} = \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{\mu}{r}\right) = -\frac{\mu}{r^2}$$

Thus

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{\mu}{2r^2}\sin\beta \tag{A6}$$

Substituting in equation (A5) gives

$$p = \frac{1}{2}\frac{\mu}{r^2}\sin\beta \tag{A7}$$

which is the same as equation (15). The remainder of the theory now proceeds as in the text.

Nomenclature

g_0 = acceleration of gravity at Earth's surface
k = $\tan\beta$ = a constant of the equi-angular spiral
m = mass of the vehicle, a variable
m_0 = mass of vehicle at start
m_p = mass of propellant ejected to achieve characteristic velocity V
p = T/m = Thrust-to-mass ratio = acceleration which the vehicle would experience if acted on by the thrust only
r = radial distance from the attracting body to the vehicle, see Fig. 1

r_0 = value of r at which spiral starts (where $\theta = 0$), see Fig. 1
r_1 = value of r after one revolution round Earth
r_c = radius of a circular orbit
r_f = value of r at which transfer of the vehicle along a spiral terminates
s = distance along flight path (see Fig. 2 for ds)
t = time (measured from instant when vehicle is at r_0)
v = speed of the vehicle
Δv = change in speed
v_c = speed of a satellite in a circular orbit
v_{c0} = value of v_c for radius equal to r_0
v_{cf} = value of v_c for radius equal to r_f
E = total energy (kinetic plus potential) of vehicle per unit mass
R = radius of curvature of spiral
T = thrust provided by vehicle's engine
U = speed, relative to vehicle, with which propellant is ejected
V = characteristic velocity for a maneuver
X = $r\omega^2$=centrifugal acceleration (as measured along radius from attracting body)
Y = externally applied forces
β = angle between flight path and a perpendicular to the radius vector, see Fig. 2
θ = angle swept through by radius vector connecting attracting center to vehicle, as vehicle proceeds along flight path, see Fig. 1
ω = $\mathrm{d}\theta/\mathrm{d}t$ = rate of rotation of radius vector connecting attracting center to vehicle
ω_c = angular velocity of satellite about center in a circular orbit
ψ = angle between tangent to flight path and a fixed line in the plane of the spiral
μ = gravitational constant for the attracting center

REFERENCES

1. A. E. Kunen and W. McIlroy, *Republ. Aviation Corp. Plasma Propulsion Lab. Rep.* No. 101 (1957).
2. M. J. Minneman, *Proc. 3rd Nat. Convention Military Electronics* p. 167 (1959).
3. S. L. Eilenberg and A. L. C. Huebner, Semi-annual meeting, American Rocket Society, June 1959, Preprint 880–59.
4. J. H. Huth, *Astronautics* p. 24 (Oct. 1958).
5. A. E. Kunen, *Republ. Aviation Corp. Plasma Propulsion Lab. Rep.* No. 108 (1958).
6. E. Stuhlinger, *J. Astronautics* **3** (1956).
7. R. Bacon, *Amer. J. Phys.* **27**, 164 (1959).
8. G. F. Forbes, *J. Brit. Interplanet. Soc.* **9**, 75 (1950).

SATELLITE ORBITS DERIVED FROM A GRAVITATIONAL MODEL OF THE EARTH

PAUL H. SAVET
Arma Division, American Bosch Arma Corporation

Abstract—In addition to disturbances caused by the residual atmosphere at satellite heights, irregularities in the orbits of so-called close satellites are caused by the finite mass distribution of the Earth. This is at variance with phenomena in usual celestial mechanics, where the perturbations created pertain to the presence of stars assimilated to point masses.

It is shown that a faithful gravitational picture of the Earth may be offered by means of a fictitious mass distribution confined to the equatorial plane of the Earth. This solution, first derived by Maxwell in a somewhat analogous electrostatic problem had been extended to gravitational phenomena and a family of actual equipotential surfaces have been derived.

The dynamical analysis resulting from this particular structure provides two prime integrals in finite terms, supplemented by another equation directly related to the period of the orbit. The latter is written in an approximate way, and the overall discussion provides a direct numerical explanation of the precession of the plane of the satellite orbit around the polar axis from east to west if the satellite motion, as projected upon the equatorial plane takes place from west to east; it also accounts for the motion of the major axis of the ellipse within the plane of the orbit; finally it explains the distortion of the orbit within its own plane.

Dramatically enough, within the last two years or so mankind managed to change the composition of the nearby universe. In these history-making accomplishments two artificial planets have been added to our solar system, while the number of moons which now supplement our old companion, faithfully circling the Earth, has been brought to an impressive magnitude. These scientific and technical triumphs are the combined result of highly developed rocketry, sophisticated guidance schemes, and all the well known supporting means put together for the attainment of a planned goal.

Yet from the point of view of the existing universe, these man-made intruders of the solar system did not change at all the established order of celestial history. Enormously larger satellites would have to be launched to cause any appreciable perturbation of the motion of celestial bodies, as governed essentially by Newton's law of gravitation. This law is referred to inertial space, in which angular orientation is determined by the motionless appearance of the overwhelming mass of the universe.

In dealing with artificial satellites, the problems are very different from those usually encountered in celestial mechanics, in which "distant" stars with respect to each other are considered, and as a result the assumption is made that their masses are concentrated in their centers of gravity. This assumption is natural enough on a truly astronomical scale, where individual stars always appear (from a mechanical point of view) as mere points. The exceptions to this rule are few, except perhaps within the solar system.

In studying satellites orbiting around the Earth, it may be legitimate to neglect the finite dimensions of these satellites, but it is not permissible to assume that the mass of the Earth is concentrated in its center of gravity. No deep insight is required to foresee, indeed, that the particular distribution of this mass within the Earth is of great enough importance to influence strongly the satellite's motion. But the converse is also true: An accurate observation of the motion of a satellite around the Earth, as compared to Kepler's laws, will provide valuable information as to the gravitational behavior of the Earth.

The mathematical problems associated with

this endeavor have been clearly formulated and a great many contributions have already been made. The purpose of this paper is neither a duplication nor a refinement of these valuable contributions; it is rather an attempt to understand a few basic facts in the realm of satellite dynamics, assuming that enough simplifications have been introduced to be able to handle mathematical discussions explicitly, while assuming that these simplifications are not far reaching enough to destroy the essential features of the phenomena to be investigated.

A better insight into the nature of satellite dynamics is provided, if a representation through an *equivalent* system is considered.

In the following (as pointed out) the only law of nature explicitly applied is Newtonian gravitation and the ensuing classical dynamics in inertial space. The latter is supposedly attached to the center of the Earth in "free fall" in the Universe and oriented with respect to remote stars. No drag of any origin is assumed, although, as is well known, it cannot be entirely neglected when the perigee height of a man-made moon is barely a small fraction of the radius of the Earth.

GRAVITATIONAL FIELD OF THE EARTH

Let us consider a satellite orbiting at a relatively low altitude, as compared to the radius of the Earth. The prevailing gravitational field cannot be considered as created by a single mass. The actual distribution of this mass is of no *explicit* importance, since it is possible (at least theoretically) to construct an *equivalent* mass distribution. The latter may be more convenient for use in calculations and yet represents closely enough the classical gravitational field of the Earth—leaving recent discoveries of a pear-shaped Earth out of consideration. In particular, if it be assumed that one of the equipotential surfaces of gravity is an oblate spheroid of axis of revolution coinciding with the polar axis of the Earth, all the other equipotential surfaces are oblate spheroids. In any meridian plane we would have a *family of equipotential confocal ellipsoids*. An appropiater mass density distribution along any particular spheroid provides the desired field intensity outside this surface. The choice of the particular spheroid for this operation is arbitrary, provided it be *within* the field. It is also possible to select a particular spheroid shrunk to a disk in the equatorial plane, and still generate the required field, provided an appropriate mass density distribution upon this disk is assumed, as will be seen later. *Ultimately we may end up with an equivalent mass (or density) distribution within the equatorial plane.*

Exploration of the gravitational field at various geographical locations has been carried out and yielded an average value, function of latitude only. Correcting for the rotation (centrifugal effect) and the oblateness of the Earth, it becomes possible to determine a *geoid*, which is one of the equipotential surfaces mentioned previously. If, according to Jeffreys,[1] a and b denote the equatorial and polar semi axes of the geoid respectively, one has an approximate oblateness

$$\epsilon = \frac{a - b}{a} \approx 1/611 \tag{1}$$

A rather simple way of deriving this value is offered, if we denote by a and b' the actual semi axes of the Earth respectively. b' is of course different from b (equal to $b' + \Delta b$). Remembering that the radius of normal curvature at the polar region of the Earth is a^2/b', one should have approximately a ratio of gravitational field strengths E_a, E_b and $E_{b'}$ respectively

$$\frac{E_b}{E_{b'}} = \left[\frac{a^2/b'}{a^2/(b' + \Delta b)}\right]^2 \cong 1 + 2\frac{b'\Delta b}{a^2} \tag{2}$$

while

$$\frac{E_a}{E_b} = \frac{a}{b} = \frac{a}{b' + \Delta b} \tag{3}$$

Magnitudes E_a, $E_{b'}$, a and b' are known, while b and E_b are to be determined from equations (2) and (3).

Taking

$$\begin{aligned} E_a &= 981{\cdot}4 \text{ cm/sec}^2 \\ E_{b'} &= 983{\cdot}2 \text{ cm/sec}^2 \\ a &= 6378 \text{ km} \\ b' &= 6357 \text{ km} \end{aligned} \tag{4}$$

we get $\Delta b \approx 11$ km (and thus $b \approx 6368$ km) and $E_b = 979{\cdot}8$ cm/sec^2. The relative oblateness would appear then as $\epsilon \approx 1/609$, as against the slightly different, more correct, value.

GRAVITATIONAL POTENTIAL

Let us use cylindrical co-ordinates (r,z,ϕ) as shown in Fig. 1. Cross-section $\phi = \text{const.}$ represents a family of equipotential surfaces of ellipsoids of revolution defined by the equation

$$\frac{r^2}{a^2 - \lambda} + \frac{z^2}{b^2 - \lambda} = 1 \tag{5}$$

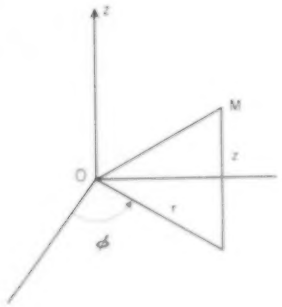


Fig 1.

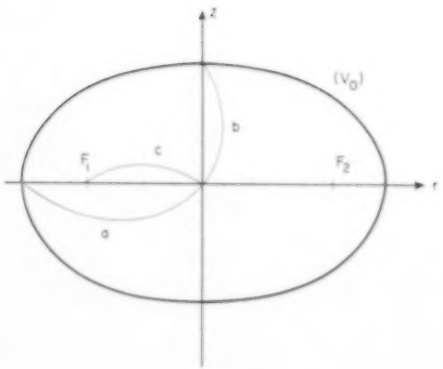


Fig. 2.

One also has $c^2 = a^2 - b^2$. Furthermore the parameter λ provides the following

$-\infty < \lambda < 0$	ellipse outside V_0 (defined by $\lambda = 0$)
$\lambda = 0$	ellipse V_0
$0 < \lambda < b^2$	confocal ellipses inside V_0
$\lambda = b^2 - 0$	segment between F_1 and F_2
$\lambda = b^2 + 0$	segment outside F_1 and F_2
$b^2 < \lambda < a^2$	confocal family of hyperbolae
$\lambda > a^2$	imaginary

(6)

Let us denote by λ_1 a parameter relative to an ellipse, while λ_2 defines a hyperbola. If the absolute field intensity of gravitation in elliptic co-ordinates (λ_1, λ_2) is $E(\lambda_1, \lambda_2)$, the expression

$$2\pi r\,(\mathrm{d}s)_{\lambda_1 - \text{const}} \times E(\lambda_1, \lambda_2) \tag{7}$$

should be an invariant with respect to λ_1, except for $\lambda_1 = b^2$, i.e. for the z axis.[2] Since

$$\lambda^2 - \lambda(a^2 + b^2 - r^2 - z^2) + \\ + a^2b^2 - (a^2z^2 + b^2r^2) = 0$$

and thus

$$\lambda_1 + \lambda_2 = a^2 + b^2 - r^2 - z^2$$

$$\lambda_1\lambda_2 = a^2b^2 - (a^2z^2 + b^2r^2) \tag{8}$$

it follows that

$$c^2r^2 = (a^2 - \lambda_1)(a^2 - \lambda_2)$$

$$c^2z^2 = -(b^2 - \lambda_1)(b^2 - \lambda_2) \tag{9}$$

On the other hand we have

$$(\mathrm{d}s)^2_{\lambda_1 - \text{const}} = \left[\left(\frac{\partial r}{\partial \lambda_2}\right)^2 + \left(\frac{\partial z}{\partial \lambda_2}\right)^2\right] \mathrm{d}\lambda_2^2 \tag{10}$$

and from equations (9)

$$\frac{2}{r}\frac{\partial r}{\partial \lambda_2} = -\frac{1}{a^2 - \lambda_2}$$

$$\frac{2}{z}\frac{\partial z}{\partial \lambda_2} = \frac{1}{b^2 - \lambda_2} \tag{11}$$

or again

$$\left(\frac{\partial r}{\partial \lambda_2}\right)^2 = \frac{1}{4c^2}\frac{a^2 - \lambda_1}{a^2 - \lambda_2}$$

$$\left(\frac{\partial z}{\partial \lambda_2}\right)^2 = -\frac{1}{4c^2}\frac{b^2 - \lambda_1}{b^2 - \lambda_2}$$

$$\left(\frac{\partial s}{\partial \lambda_2}\right)^2 = \frac{1}{4}\frac{\lambda_1 - \lambda_2}{(a^2 - \lambda_2)(b^2 - \lambda_2)} \tag{12}$$

Writing the condition that

$$r\frac{\partial s}{\partial \lambda_2} E(\lambda_1, \lambda_2)$$

is a function of λ_2 alone, mentioned equations provide

$$\frac{1}{2c}\sqrt{\left[\frac{a^2-\lambda_1}{b^2-\lambda_2}(\lambda_1-\lambda_2)\right]}\times E(\lambda_1,\lambda_2)=\phi(\lambda_2) \tag{13}$$

For $\lambda_1 \approx -\infty$ we have

$$E=2c\,\phi(\lambda_2)\sqrt{\left(-\frac{b^2-\lambda_2}{\lambda_1^2}\right)} \tag{13a}$$

and this should be equal to Q_0/λ_1 in order to satisfy the conditions at infinity, while supposing that the potential is due to an equivalent (fictitious) distribution of masses in the equatorial plane $z=0$.

This mass is distributed between $0<r<c$ in a total amount equal to Q_0. It follows that

$$\phi(\lambda_2)=-\frac{Q_0}{2c\sqrt{[-(b^2-\lambda_2)]}} \tag{14}$$

and therefore

$$E=-\frac{Q_0}{\sqrt{[(a^2-\lambda_1)(\lambda_2-\lambda_1)]}} \tag{15}$$

Exchanging λ_1 and λ_2 in equation (12) offers

$$\left(\frac{\partial s}{\partial\lambda_1}\right)^2_{\lambda_2=\text{const}}=\frac{1}{4}\frac{\lambda_2-\lambda_1}{(a^2-\lambda_1)(b^2-\lambda_1)} \tag{16}$$

and since $E=-\operatorname{grad} V$, where V is the potential function,

$$E=-\frac{\partial V}{\partial\lambda_1}\left(\frac{d\lambda_1}{ds}\right)=-\frac{Q_0}{\sqrt{[(a^2-\lambda_1)(\lambda_2-\lambda_1)]}}$$

$$=-2\frac{\partial V}{\partial\lambda_1}\sqrt{\left[\frac{(a^2-\lambda_1)(b^2-\lambda_1)}{\lambda_2-\lambda_1}\right]} \tag{17}$$

or

$$\frac{\partial V}{\partial\lambda_1}=\frac{Q_0}{2(a^2-\lambda_1)\sqrt{(b^2-\lambda_1)}}$$

Thus

$$V=\int_{-\infty}^{\lambda_1}\frac{Q_0\,d\lambda_1}{2(a^2-\lambda_1)\sqrt{(b^2-\lambda_1)}}$$

Remembering the formula

$$\int\frac{dx}{(a^2-x)\sqrt{(b^2-x)}}=-\frac{2}{c}\arctan\frac{\sqrt{(b^2-x)}}{c}+\text{const}$$

we have

$$V=-\frac{Q_0}{c}\left[\arctan\frac{\sqrt{(b^2-\lambda_1)}}{c}\right]_{-\infty}^{\lambda_1}$$

$$=\frac{Q_0}{c}\operatorname{arc\,cot}\frac{\sqrt{(b^2-\lambda_1)}}{c} \tag{18}$$

The surface density (σ) of distribution of matter within the equatorial plane $z=0$, i.e. for $\lambda_1=b^2$ and $r^2=a^2-\lambda_2$ is given by Poisson's equation, expressed as

$$\sigma=\frac{1}{2\pi}E_{(\lambda_1=b^2)}=\frac{Q_0}{2\pi c\sqrt{[-(b^2-\lambda_2)]}}$$

$$=\frac{Q_0}{2\pi c\sqrt{(c^2-r^2)}} \tag{19}$$

The total amount of distributed matter in the equatorial plane ($z=0$) is

$$\int_0^c 2\pi\sigma dr=\int_0^c\frac{Q_0\,r\,dr}{c\sqrt{(c^2-r^2)}}=Q_0 \tag{20}$$

as anticipated.[3]

Having established the gravitational potential of an equivalent mass distribution in the equatorial plane ($z=0$), a natural simplification can be introduced by remembering that $\epsilon=(a-b)/a$ is small, and therefore higher powers of ϵ can be neglected.

Let us introduce in the $\phi=\text{const}$ plane (see Fig. 3) a polar angle u, which is essentially a geocentric latitude, together with an eccentric anomaly (ζ) and a true geographical latitude (η). Magnitude R represents the distance from the center of the Earth (OM).

By definition

$$\frac{r}{\sqrt{(a^2-\lambda)}}=\cos\zeta$$

$$\frac{z}{\sqrt{(b^2-\lambda)}}=\sin\zeta \tag{21}$$

while $\lambda=a^2\cos^2\zeta+b^2\sin^2\zeta-R^2$. Since ϵ is small, calculations yield within first orders of magnitude

$$\cos\zeta=\cos u\left(1-\frac{a^2\sin^2 u}{a^2-\lambda}\,\epsilon\right)$$

$$\sin \zeta = \sin u \left(1 + \frac{a^2 \cos^2 u}{a^2 - \lambda} \epsilon\right)$$

$$\lambda = a^2 (1 - 2 \sin^2 u \, \epsilon) - R^2 \qquad (22)$$

It follows that

$$a^2 - \lambda = R^2 + 2a^2 \sin^2 u \, \epsilon$$

$$b^2 - \lambda = R^2 - 2a^2 \cos^2 u \, \epsilon$$

$$c^2 = 2a^2 \epsilon \qquad (23)$$

and also, in virtue of equation (22),

$$\cos \zeta = \cos u \left(1 - \frac{a^2}{R^2} \sin^2 u \, \epsilon\right)$$

$$\sin \zeta = \sin u \left(1 + \frac{a^2}{R^2} \cos^2 u \, \epsilon\right) \qquad (24)$$

expressions limited to the first power of ϵ. Based on equations (21)–(23), we have in the $\phi = \text{const}$ plane

$$\begin{matrix} r = R \cos u \\ z = R \sin u \end{matrix} + \text{terms in } \epsilon^2 \qquad (26)$$

Also, the absolute velocity squared of a point satellite in space appears as

$$\left(\frac{ds}{dt}\right)^2 = \left(\frac{dR}{dt}\right)^2 + R^2 \left(\frac{du}{dt}\right)^2 + R^2 \cos^2 u \left(\frac{d\phi}{dt}\right)^2 \qquad (27)$$

A first (and rather obvious) prime integral of the motion is offered by remembering that the potential does not depend on ϕ. Since the

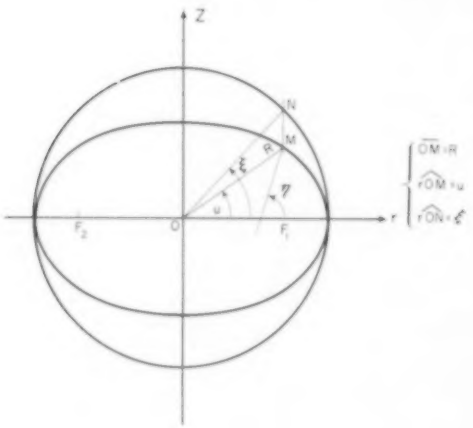


Fig. 3.

The gravitational potential, as previously derived, can now be written

$$V = \frac{Q_0}{c} \operatorname{arc\,cot} \frac{\sqrt{(b^2 - \lambda_1)}}{c}$$

$$= \frac{Q_0}{a\sqrt{(2\epsilon)}} \operatorname{arc\,cot} \frac{R\,[1 - (a^2/R^2) \cos^2 u \, \epsilon]}{a\sqrt{(2\epsilon)}}$$

$$= -\frac{Q_0}{R} \left[1 + \left(\frac{a^2}{R^2} \cos^2 u - \frac{2}{3}\right) \epsilon + \ldots\right] \qquad (25)$$

EQUATIONS OF MOTION

We shall use spherical co-ordinates (R, u, ϕ) as shown in Fig. 4, and rely on approximate

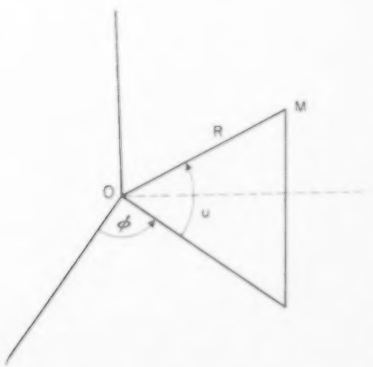


Fig. 4.

gravitational field vector always intersects the polar axis of the Earth (without generally passing through its center O), one has the projected area rule upon the equatorial plane

$$R^2 \cos^2 u \frac{d\phi}{dt} = h \tag{28}$$

A second prime integral is derived by the application of the theorem of conservation of energy in a non-dissipative system, viz.

$$\frac{1}{2}\left(\frac{ds}{dt}\right)^2 + V = k \tag{29}$$

where k is another constant, depending on the initial conditions, while the other magnitudes are those appearing in equations (25) and (27).

Since a third prime integral does not appear in an obvious way, let us write the force vs. acceleration equation for the variable u. The acceleration relative to u is expressed by

$$J_u = \frac{1}{R}\frac{d}{dt}\left(R^2 \frac{du}{dt}\right) + R \sin u \cos u \left(\frac{d\phi}{dt}\right)^2 \tag{30}$$

as can be derived by considering the equation of the velocity in vector notations (i.e. algebraic value of the velocity, multiplied by an appropriately oriented unit-vector) and remembering the derivatives of the unit vector with respect to u and ϕ. Above expression of J_u should be equal to

$$J_u = -\frac{1}{R}\frac{\partial V}{\partial u} = -\frac{Q_0 a^2}{R^4} \sin 2u \, \epsilon \tag{31}$$

or explicitly

$$\frac{d}{dt}\left(R^2 \frac{du}{dt}\right) + R^2 \sin u \cos u \left(\frac{d\phi}{dt}\right)^2 = -\frac{Q_0 a^2}{R^3} \sin 2u \, \epsilon \tag{32}$$

The time derivative of ϕ (i.e. $d\phi/dt$) can be eliminated by using formula (28). Equations (27), (29) and (32) yield

$$\frac{1}{2}\left[\left(\frac{dR}{dt}\right)^2 + R^2\left(\frac{du}{dt}\right)^2 + \frac{h^2}{R^2 \cos^2 u}\right] - \frac{Q_0}{R}\left[1 + \frac{a^2}{R^2}\left(\cos^2 u - \frac{2}{3}\right)\epsilon\right] = k$$

$$\frac{d}{dt}\left(R^2 \frac{du}{dt}\right) + \frac{h^2 \sin u}{R^2 \cos^3 u} = -\frac{Q_0 a^2}{R^3} \sin 2u \, \epsilon$$

$$\frac{d\phi}{dt} = \frac{h}{R^2 \cos^2 u} \tag{33}$$

The first two equations provide R and u, while the third one offers ϕ as a function of R and u. The complete integration of system (33) requires the use of a computing machine. It is possible, however, to derive interesting conclusions, if not a complete solution, by directly analysing these equations.

First, let us take two particular cases, viz. $u = 0$ and $\phi = 0$.

Motion within the Equatorial Plane ($u = 0$)

The second equation in (33) vanishes. First and third equations become respectively

$$\left(\frac{dR}{dt}\right)^2 = 2k + \frac{2Q_0}{R} - \frac{h^2}{R^2} + \frac{2}{3}\frac{Q_0 a^2}{R^3}\epsilon$$

$$\frac{d\phi}{dt} = \frac{h}{R^2} \tag{34}$$

For $\epsilon = 0$ we recognize equations (34) as representing a conical section, the nature of which depends on the sign of k. If this is positive, we have a hyperbola; for $k < 0$ we get an ellipse; while $k = 0$ corresponds to a parabola.

Let us take the case of an *ellipse* and consider again equations (34). If ϵ is different from zero, curve $(dR/dt)^2$ vs. $1/R$ undergoes a slight distortion as shown in Fig. 5. Since $(dR/dt)^2$ has to be positive, the fact that ϵ is positive merely alters the extreme values of R between which a real solution of equation (34) exists. Furthermore $\epsilon > 0$ suggests that $R(\phi)$ should get expressed in terms of elliptic rather than trigonometric functions, so that a slight shift in ϕ may occur after every revolution along the orbit. We are faced with a precession of the orbit as a whole in the equatorial plane.

In order to get a more precise relationship between the precession and the basic orbit parameters, let us write again equation (34), after the substitutions

$$\frac{1}{R} = P$$

$$N = 2k + 2Q_0 P - h^2 P^2 \tag{35}$$

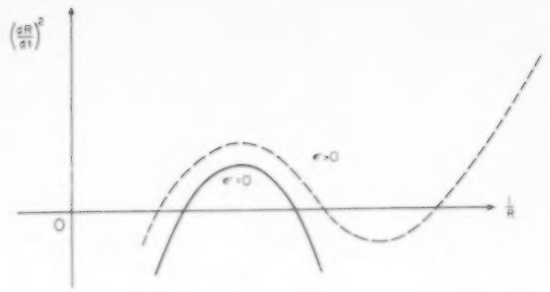


Fig. 5.

which provide within the first power of ϵ

$$\frac{dP}{d\phi} = \frac{1}{h}\sqrt{N}\left(1 + \frac{1}{3}\frac{Q_0 a^2 P^3}{N}\epsilon\right) \tag{36}$$

For $\epsilon = 0$, P would appear as a sine curve (Fig. 6).

At M, the slope $dP/d\phi$ is a maximum, which occurs at $N = \max.$ or

$$P = P_M = \frac{Q_0}{h^2} \tag{37}$$

with

$$N_{\max} = 2k + \left(\frac{Q_0}{h}\right)^2 \tag{37a}$$

On the other hand, if ϵ^* denotes the eccentricity of the *orbit*, we have

$$k = -\frac{1}{2}\left(\frac{Q_0}{h}\right)^2 (1 - \epsilon^*)^2 \tag{37b}$$

As a result

$$N_{\max} = \left(\frac{Q_0}{h}\right)^2 [1 - (1 - \epsilon^*)^2] \tag{38}$$

An *expansion* in $\Delta\phi$ (i.e. the half angular period) is introduced as a result of changing the slope of $P(\phi)$ in M. Specifically, one has

$$\frac{\delta(dP/d\phi)}{dP/d\phi} \approx -\frac{\delta(\Delta\phi)}{\Delta\phi} = \frac{1}{3}\frac{Q_0 a^2 P_M^3}{N_{\max}}\epsilon =$$

$$= \frac{1}{3}\frac{Q_0^2 a^2}{h^4}\frac{\epsilon}{1 - (1 - \epsilon^*)^2} \tag{38a}$$

As an order of magnitude, for a low altitude almost circular orbit, we have

$$Q_0 \approx gR^2$$
$$a \approx R$$
$$h = \sqrt{(gR^3)}$$

and therefore

$$-\frac{\delta(\Delta\phi)}{\Delta\phi} \approx \frac{1}{3}\frac{\epsilon}{1 - (1 - \epsilon^*)^2} \tag{39}$$

This approximation is valid only of $\epsilon^* \neq 0$, or more precisely if $\frac{1}{3}\frac{\epsilon}{1 - (1 - \epsilon^*)^2} \ll 1$, or again

$$\epsilon^* \gg \frac{\epsilon}{6} \tag{40}$$

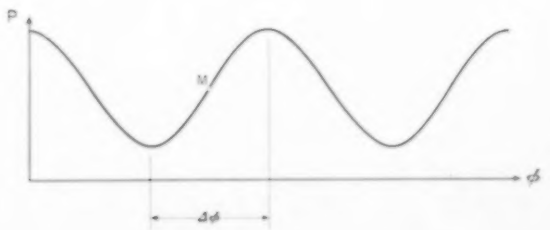


Fig. 6.

As the sign of $\delta(\Delta\phi)$ is negative, the precession is *opposite* to the direction in which the orbit is described.

Motion within an Inertial Meridian Plane ($\phi = 0$)

The last one of equations (33) vanishes on account of $h = 0$, while the second equation, after multiplication with $R^2\,\mathrm{d}u/\mathrm{d}t$ provides

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(R^2\frac{\mathrm{d}u}{\mathrm{d}t}\right)^2 = -\frac{Q_0 a^2}{R}\sin 2u\,\epsilon\frac{\mathrm{d}u}{\mathrm{d}t} \quad (41)$$

Let us assume, as an approximation, that R in the second member be considered constant. An explicit, though approximate, integration of equation (41) becomes possible, and this operation successively yields

$$\left(R^2\frac{\mathrm{d}u}{\mathrm{d}t}\right)^2 \approx h_1^2 + \frac{Q_0 a^2}{R}\cos 2u\,\epsilon \quad (42)$$

where h_1 is a constant of integration. Extraction of the square root results in

$$R^2\frac{\mathrm{d}u}{\mathrm{d}t} \approx h_1\left(1 + \frac{Q_0 a^2}{2h_1^2 R}\cos 2u\,\epsilon\right) \quad (42a)$$

Combination with the first of equations (33) yields

$$\left(\frac{\mathrm{d}R}{\mathrm{d}t}\right)^2 = 2k + \frac{2Q_0}{R} - \frac{h_1^2}{R^2} - \frac{Q_0 a^2}{3R^3}\,\epsilon \quad (43)$$

or finally after eliminating $\mathrm{d}t$ from equations (42) and (43)

$$\left(\frac{\mathrm{d}R}{\mathrm{d}u}\right)^2 = \frac{N_1\left(1 - \dfrac{Q_0 a^2}{3R^3 N_1}\,\epsilon\right)}{\dfrac{h_1^2}{R^4}\left(1 + \dfrac{Q_0 a^2}{R h_1^2}\cos 2u\,\epsilon\right)}$$

$$N_1 = 2k + \frac{2Q_0}{R} - \frac{h_1^2}{R^2} \quad (44)$$

Extracting the square root and putting $P = 1/R$ as before, we also have

$$h_1\frac{\mathrm{d}P}{\mathrm{d}u} \approx \sqrt{N_1}\left(1 - \frac{Q_0 a^2 P^3\epsilon}{6N_1} - \frac{Q_0 a^2 P}{2h_1^2}\cos 2u\,\epsilon\right) \quad (45)$$

Proceeding along the trend of thought used in the previous case ($u = 0$), and denoting by ϵ_1^* the eccentricity of *this* particular orbit, we have at the point M of Fig. 7

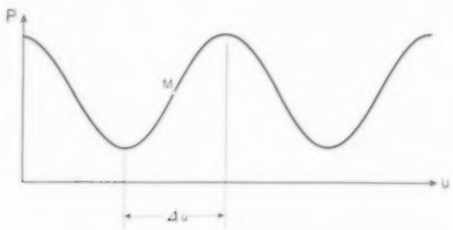


Fig. 7.

$$P_M = \frac{Q_0}{h_1^2}$$

$$(N_1)_M = \left(\frac{Q_0}{h_1}\right)^2\left[1 - (1 - \epsilon_1^*)^2\right] \quad (46)$$

and as a result, according to the reasoning followed in mentioned previous case

$$\frac{\delta(\Delta u)}{\Delta u} = \frac{Q_0^2 a^2}{h_1^4}\left(\frac{1}{6\,[1 - (1 - \epsilon_1^*)^2]} - \frac{1}{2}\cos 2u\right)\epsilon \quad (47)$$

If the eccentricity of the orbit is very small (specifically if $\epsilon_1^* < 1 - \sqrt{(2/3)} \approx 0{\cdot}2$), the first

term within the parenthesis of the second member prevails over the second, which shows that $\delta(\Delta u) > 0$, i.e. the orbit precesses in the direction of its being described, in addition to an oscillatory precession proportional to cos $2u$, which manifests itself as a distortion, due to the departure of any equipotential surface from a spherical shape. If this *distortion* be disregarded, we have

$$\frac{\delta(\Delta u)}{\Delta u} \approx \frac{Q_0^2 a^2}{6h_1^3} \frac{\epsilon}{1-(1-\epsilon_1^*)^2} > 0 \qquad (48)$$

as a true (though approximate) expression of the relative precession of the orbit.

Motion of a General Nature

The discussion of this case is more tedious than the previous two particular motions. No definite attempt is made here to follow, all the way through, the intuitive expectation that one should witness a particular *combination* of the results previously derived. That this expectation is well founded, however, appears by considering the second equation (33), with the approximation already mentioned, as far as the second member is concerned.

Denoting by H a constant of integration, we have successively

$$\left(R^2\frac{du}{dt}\right)^2 + \frac{h^2}{\cos^2 u} \approx \frac{Q_0 a^2}{R}\cos 2u\,\epsilon + H^2$$

$$\left(\frac{dR}{dt}\right)^2 = 2k + \frac{2Q_0}{R} - \frac{H^2}{R^2} - \frac{Q_0 a^2}{3R^3}\epsilon$$

$$\frac{d\phi}{dt} = \frac{h}{R^2\cos^2 u} \qquad (49)$$

Eliminating the time from the first two equations, while putting as before

$$\frac{1}{R} = P$$

$$N_2 = 2k + 2Q_0 P - H^2 P^2 \qquad (50)$$

we have

$$\sqrt{\left(H^2 - \frac{h^2}{\cos^2 u}\right)}\frac{dP}{du} = \sqrt{N_2}$$

$$\left\{1 - \frac{Q_0 a^2 P^3 \epsilon}{6N_2} - \frac{Q_0 a^2 P \cos 2u}{2\left[H^2 - (h^2/\cos^2 u)\right]}\epsilon\right\} \qquad (51)$$

First $H^2 > h^2/\cos^2 u$, as appears from equations (49); also $|u| \neq \pi/2$; therefore $h^2/\cos^2 u$ stays finite.

The resemblance with equation (45) is striking, except for the occurrence of $H^2 - h^2/\cos^2 u$ in the denominator of the second term in the parenthesis. As pointed out previously, the first term in the parenthesis contributes to the steady precession, while the cos $2u$ term brings about a distortion of the orbit.

As far as the relationship between R and ϕ is concerned, we have

$$\frac{dP}{d\phi} \approx \frac{\cos^2 u}{h}\sqrt{N_2}\left(1 - \frac{Q_0 a^2 P^3}{6N_2}\epsilon\right) \qquad (52)$$

the interpretation of which is very similar to corresponding relationships established previously.

CONCLUSIONS

Foregoing analysis reveals that the particular mass distribution of the Earth causes satellite orbits to precess in a direction and magnitude which appears to be verified by observed data. These phenomena, together with the distortion of orbits are caused by the shape of the geoid, and the latter may be represented by an *equivalent* mass distribution which lends itself to the direct analysis of satellite dynamics.

Nomenclature

a	equatorial semi axis of the Earth
b'	polar semi axis of the Earth
b	polar semi axis of geoid (equatorial axis = a)
E_a	gravity field at the equator
$E_{b'}$	same at the pole of the Earth
E_b	same at the pole of the geoid
Q_0	scale factor in the potential function
ϵ	relative eccentricity of geoid
ϵ^*, ϵ_1^*	eccentricity of orbits
c	$\sqrt{(a^2 - b^2)}$ relative to geoid
λ, λ_1, λ_2	parameters in equation of potential

r distance from the polar axis
z distance along the polar axis
E field intensity
s length of arc
V potential function
σ surface density of mass in equatorial plane
ζ eccentric anomaly
u geocentric latitude
R distance from the center of the Earth
N see equation (35), etc.
N_1 see equation (44), etc.
N_2 see equation (50), etc.
ϕ meridian angle
h, h_1, H, k constants of integration
J_u acceleration for variable u
P $1/R$

REFERENCES

1. H. Jeffreys, *The Earth*, Cambridge University Press (1952).
2. H. Bateman, *Partial Differential Equations of Mathematical Physics* pp. 421–425, Dover Publ. (1944).
3. J. Jeans, *Mathematical Theory of Electricity and Magnetism* p. 249, Cambridge University Press (1941). J. C. Maxwell, *A Treatise on Electricity and Magnetism* Vol. 1, Chap. X, (1881), Dover, New York (1954).

THE BI-ELLIPTICAL TRANSFER BETWEEN CO-PLANAR CIRCULAR ORBITS

RUDOLF F. HOELKER and ROBERT SILBER
Army Ballistic Missile Agency, Aeroballistics Laboratory*

Abstract—An investigation of the area of relative altitudes of co-planar circular orbits for which a bi-elliptical transfer is more economical than the classical Hohmann transfer. It is found that for transfers between orbits apart by a radii-ratio larger than approximately 15·58, the Hohmann transfer is less economic than any bi-elliptical transfer with an intermediate point of altitude larger than that of the larger terminal orbit. For radii-ratios less than 15·58 but greater than approximately 11·94, the bi-elliptical transfer is more economical if the intermediate point is placed at a sufficiently high altitude. This altitude is infinite at 11·94 and approaches the altitude of the larger terminal orbit at 15·58.

INTRODUCTION

In 1925, Hohmann published a book[1] in which he suggested a method for transferring (a vehicle) from a circular orbit to a co-planar (concentric) circular orbit of different radius. This method, which has come to be called the Hohmann ellipse or transfer ellipse, consists of a half-ellipse, originating and terminating at opposite apsides. We shall consider herein that the transfer takes place from a circle of smaller altitude to one of larger, but it should be pointed out that all equations and results are equally applicable to the reverse situation, i.e. transfers from a circular orbit to one of smaller altitude or radius. The Hohmann ellipse is shown in Fig. 1. Two circular co-planar orbits of radius r_1 and r_2 are assumed about a common attraction. The attracting body is idealized insofar as it is assumed that the attracting force depends only on the distance to the center of the attracting body, as in the case of a point mass, or homogeneous sphere. Also, the mass of the orbiting vehicle is assumed to be negligible compared to the mass of the attracting body so that the attracting body may be considered stationary.

The vehicle is traveling on the inner circle in possession of a circular velocity which is a function only of the radius r_1. At some point on the orbit, labeled the starting point, a velocity increment is imparted to the vehicle causing the orbit to become elliptical. This velocity increment is assumed to be imparted instantaneously by an impulsive thrust and is co-linear with the velocity vector before application, so that the resultant vector is tangent to the path of both the circle and the ellipse and hence perpendicular to the local radius vector. Because of this, the

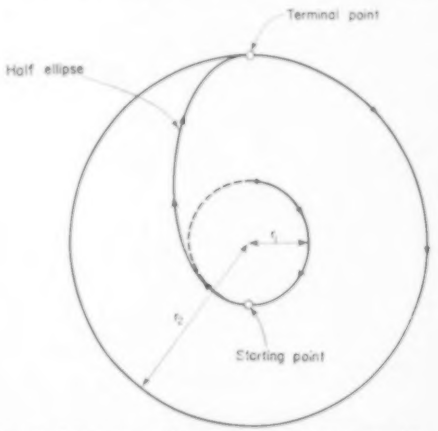


Fig. 1. Schematic of the two-point Hohmann ellipse.

starting point is at an apsis of the transfer ellipse, in this case, the perigee. The magnitude of the velocity increment is adjusted so that the opposite apsis of the transfer ellipse lies at the radius of the desired outer circle, r_2. At this point, a second velocity increment is added, again co-linear with the arrival velocity vector, and is of the magnitude necessary to yield the velocity vector which places the vehicle into a circular orbit of radius r_2. If it is assumed that

* Now — Marshall Space Flight Center, National Aeronautics and Space Administration.

the application of the two impulses is a single interrupted stage rather than a multi-stage affair, then the propellant consumption required to bring about the complete transfer is measurable by the sum of the two velocity increments, provided the latter are considered in their absolute value. Lawden[2] has shown that the Hohmann transfer, as we have just considered it, offers an optimum with respect to propellant considerations compared to all transfers employing two impulses. The question still remains, however, as to the possibility of devising a scheme of transfer involving more than two impulses which might be more economical than the Hohmann transfer. The bi-elliptical transfer is such a scheme.

PRELIMINARY INVESTIGATIONS

The two velocity increments employed in the Hohmann transfer are functions of the radii r_1 and r_2 as well as the attracting mass. If the two increments are properly summed and divided by the circular velocity of the initial orbit, the resulting expression is a function only of the *ratio* of the radii of the two considered orbits. For this reason it will be convenient henceforth to refer to the two considered orbits as the *terminal circles* and the ratio of their radii, taken in such a way as to be greater than or equal to unity, as the *terminal ratio*. The total required velocity increment, after being divided by circular velocity of the smaller terminal circle, is a dimensionless quantity and may be referred to as the *transfer modulus*. If a circular velocity of the initial (smaller) terminal circle (of radius r_1) is denoted by c_1 and the two velocity increments by Δv_1 and Δv_2, the transfer modulus may be expressed in terms of the terminal ratio r_2/r_1 by

$$\frac{\Delta v_t}{c_1} = \frac{|\Delta v_1| + |\Delta v_2|}{c_1} = \left(1 - \frac{1}{r_2/r_1}\right) \sqrt{\left[\frac{2\,(r_2/r_1)}{1 + (r_2/r_1)}\right]} + \frac{1}{\sqrt{(r_2/r_1)}} - 1 . \qquad (1)$$

The plot of equation (1) may be seen in Fig. 2. The transfer modulus is seen to reach a maximum at the circled point which corresponds to a terminal ratio of approximately 15·58. Beyond this, the transfer modulus asymptotically approaches the value $\sqrt{2} - 1$ as the terminal ratio becomes infinite. Considered physically, the quantity $(\sqrt{2} - 1)\, c_1$ is the difference in escape velocity and circular velocity for the smaller terminal circle. Hence it is seen that for some values of the terminal ratio, the Hohmann transfer modulus actually exceeds requirements for escape.

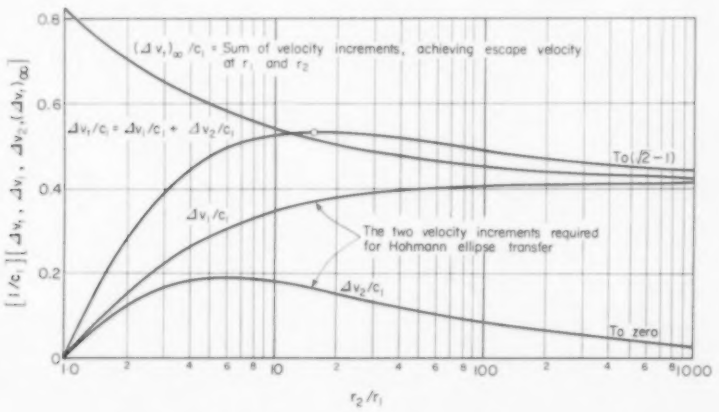

$(\Delta v_t)_\infty / c_1$ = Sum of velocity increments, achieving escape velocity at r_1 and r_2
$\Delta v_t / c_1 = \Delta v_1 / c_1 + \Delta v_2 / c_1$
To $(\sqrt{2} - 1)$
$\Delta v_1 / c_1$
The two velocity increments required for Hohmann ellipse transfer
$\Delta v_2 / c_1$
To zero
$[1/c_1]\,[\Delta v_t\ ,\ \Delta v_1\ ,\ \Delta v_2\ ,\ (\Delta v_t)_\infty]$
r_2/r_1

Fig. 2. Velocity increments for Hohmann ellipse transfer.

Because for some terminal ratios the transfer

modulus is larger than the transfer modulus for infinite terminal ratios, the following question becomes of interest. What would be the transfer modulus for some finite terminal ratio if, instead of employing a Hohmann transfer, the vehicle is sent on a Hohmann ellipse to infinity (parabola) and returned in like manner to the desired terminal orbit where an appropriate velocity increment is given to obtain circular velocity for that circle? This would mean two half-ellipses (in this case parabolas) with a common apsis at some chronologically intermediate point, in this case at infinity. The velocity requirement for this maneuver would be equal to the difference in escape velocity and circular velocity for the terminal circles added in absolute value, since no impulses would be applied at infinity. The transfer modulus for this is shown in Fig. 2 and the equation is as follows:

$$\frac{(\Delta v_t)_\infty}{c_1} = (\sqrt{2} - 1)\left[1 + \frac{1}{\sqrt{(r_2/r_1)}}\right] \qquad (2)$$

As can be seen from Fig. 2, the transfer modulus for this maneuver is equal to the Hohmann transfer modulus for a terminal ratio of approximately 11·94, and for greater terminal ratios the transfer modulus is consistently less than that for the Hohmann transfer. Both, however, approach the same value at infinite transfer ratios.

The chronologically intermediate point at infinity just considered may be generalized to any point of finite radius. The first half-ellipse would then have apsides at r_1 and at the intermediate point and the second at the intermediate point and r_2. However, this sequence of maneuvers would suggest a re-designation of radii symbols. The first point, being on the first circular orbit, logically possesses a radius of r_1, but the intermediate point is now the second point. Henceforth, the radius of the intermediate point will be referred to as r_2 and the radius of the desired (outer) terminal circle will be denoted by r_3. The terminal ratio now becomes r_3/r_1 and it will be convenient to establish at this point a term for the ratio r_2/r_1. Since the point of radius r_2 is the point of conjunction of the two half-ellipses, r_2/r_1 will be referred to as the *conjunction ratio*.

In general, the radius of the intermediate (conjunction) point, besides being finite, may be either larger or smaller than the radii of the terminal circles or in between. Two situations are shown in Fig. 3, one in which r_2 is greater than both r_1 and r_3, and one in which r_2 is greater than r_1 but less than r_3. A third situation exists, for which r_2 is less than both r_1 and r_3, and all three of these situations are to be investigated.

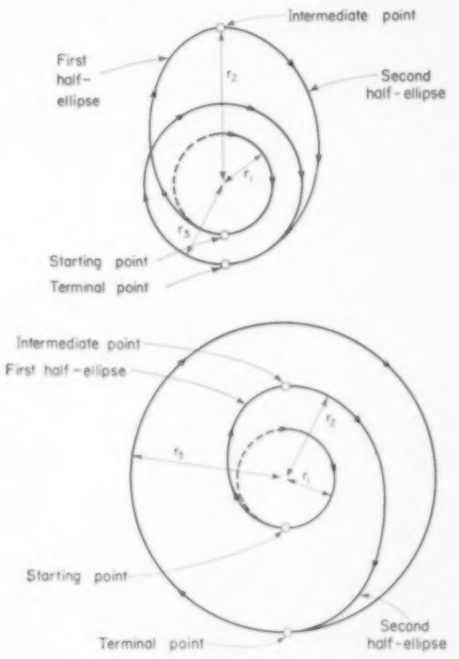


Fig. 3. Schematics of the bi-elliptical or three-point transfer.
Top: $r_2/r_1 > r_3/r_1 > 1$
Bottom: $r_3/r_1 > r_2/r_1 > 1$

The geometry of the maneuver with an intermediate point as seen in Fig. 3 suggests a natural name for this maneuver, a *bi-elliptical transfer*, composed of two half-ellipses junctioned at a common apsis. Also, since the bi-elliptical transfer involves three points of impulse application (an impulse application must be given at the intermediate point if it is

not at an infinite radius) it may alternatively be referred to as a three-point transfer as opposed to the two-point Hohmann transfer.

Having established the bi-elliptical transfer, the problem may be formulated as the establishment of what initial orbital conditions (terminal ratios) possess one or more bi-elliptical transfer moduli smaller than the Hohmann transfer modulus.

ANALYSIS OF THE BI-ELLIPTICAL TRANSFER

There are immediately distinguishable three types of bi-elliptical transfers, as has already been mentioned. Bi-elliptical transfers with conjunction ratios greater than the terminal ratio may be referred to as bi-elliptical transfers of *exterior conjunction*, since the intermediate point lies outside of both of the terminal circles. The other two cases, for which the intermediate point lies within one or both of the terminal circles may be called *interior conjunctions*. It will be seen that exterior conjunctions offer lower transfer moduli than the Hohmann transfer for certain terminal ratios, but that all interior conjunctions possess transfer moduli higher than that of the corresponding Hohmann ellipse.

As has been previously discussed, in deriving the transfer modulus, the individual velocity increments must be added in absolute value. The expression for the transfer modulus for the three types of bi-elliptical transfers may be written in a single form containing absolute values. The positive orientation of the elements within the absolute values differs for the three cases. The general expression is

$$\frac{\Delta v_t}{c_1} = \left| \sqrt{\left[\frac{2\,(r_2/r_1)}{1+(r_2/r_1)}\right]} - 1 \right| + \sqrt{\left(\frac{2}{r_2/r_1}\right)} \left| \sqrt{\left[\frac{1}{1+(r_2/r_1)/(r_3/r_1)}\right]} - \sqrt{\left(\frac{1}{1+r_2/r_1}\right)} \right| + \sqrt{\left(\frac{1}{r_3/r_1}\right)} \left| \sqrt{\left(\frac{2\,r_2/r_1}{r_3/r_1+r_2/r_1}\right)} - 1 \right| . \quad (3)$$

For bi-elliptical transfers with exterior conjunction, the positive orientation of th quantities within the absolute value signs is readily determined to be

$$\frac{\Delta v_t}{c_1} = \sqrt{\left[\frac{2\,(r_2/r_1)}{1+(r_2/r_1)}\right]} - 1 + \sqrt{\left(\frac{2}{r_2/r_1}\right)} \left[\sqrt{\left[\frac{1}{1+(r_2/r_1)/(r_3/r_1)}\right]} - \sqrt{\left(\frac{1}{1+r_2/r_1}\right)} \right] + \sqrt{\left(\frac{1}{r_3/r_1}\right)} \left\{ \sqrt{\left[\frac{2\,(r_2/r_1)}{r_2/r_1+r_3/r_1}\right]} - 1 \right\} . \quad (4)$$

Interior conjunctions will not be considered until exterior conjunctions have been thoroughly considered.

Equation (4) shows the transfer modulus for exterior conjunction to be a function only of the terminal ratio and the conjunction ratio. Fig. 4 shows the transfer modulus as a function of the conjunction ratio with the terminal ratio as a parameter. Investigation into the behavior of these curves will provide the information necessary to define the situations for which the bi-elliptical transfer is more economical than the Hohmann transfer (for exterior conjunctions).

The first point of interest for any one of the parametric curves is the derivative or slope of the curve at the first point of the curve. The first point of each of the curves is seen to lie on the broken curve which represents the transfer modulus for the Hohmann transfer. This is the point for which the terminal ratio is equal to the conjunction ratio, or, where the intermediate point lies on the larger terminal circle. As the parametric curves continue through increasing conjunction ratios, the intermediate point is moving out toward infinity yielding exterior conjunctions. It is important to realize that each parametric curve represents transfer moduli for the same initial problem—the parameter of the curve is the terminal ratio. As an example, in Fig. 4, examine the curve corresponding to a terminal ratio of 100. The initial point of the curve corresponds to a conjunction ratio of 100 and the transfer modulus at that point is the Hohmann transfer modulus. Notice that as the curve progresses through increasing conjunction ratios, the transfer modulus diminishes from the Hohmann value, indicating that this bi-elliptical transfer is more economical than the Hohmann

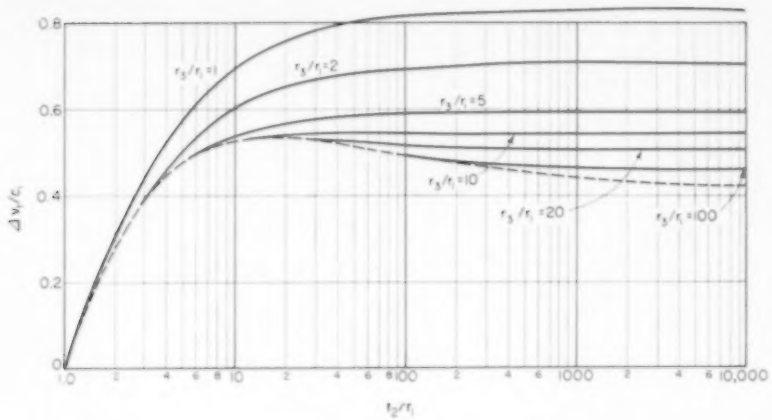


Fig. 4. Total velocity increment (expressed in multiples of c_1) necessary to effect a transfer from one circular orbit to another, the relative positions of which are given by r_3/r_1, by means of a bi-elliptical transfer, the intermediate point of which is given by r_2/r_1 vs. r_2/r_1 for various values of r_3/r_1. Broken curve indicates locus of $r_2/r_1 = r_3/r_1$, the standard Hohmann ellipse.

transfer. In fact, it would follow that *every* parametric curve which initiated with a negative slope would possess some points for which the bi-elliptical transfer modulus was less than the Hohmann value.

The slope at the initial point is obtained from differentiation of equation (4) with respect to the conjunction ratio and subsequent equating of the conjunction ratio to the terminal ratio. Hence the slope at the initial point is a function

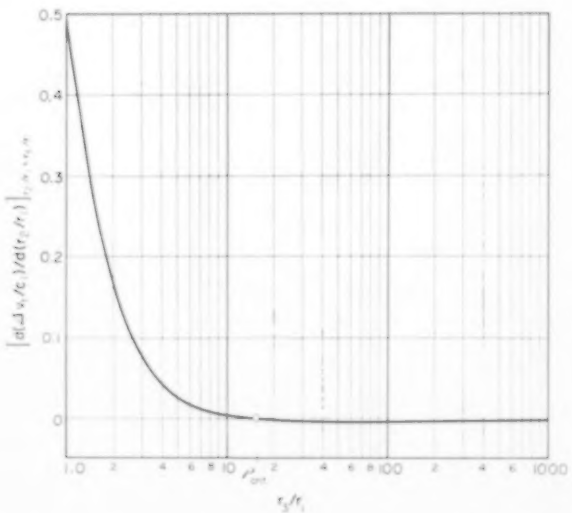


Fig. 5. Slope of total velocity increment over change of r_2/r_1, plotted versus r_3/r_1 for $r_2 = r_3$ and case of intermediate point larger than terminal orbits.

only of the terminal ratio and is shown as such in Fig. 5. The equation of this is

$$\left[\frac{\mathrm{d}\,(\Delta v_t/c_1)}{\mathrm{d}\,(r_2/r_1)}\right]_{r_2/r_1 = r_3/r_1} = \frac{\sqrt{2}}{(r_3/r_1)^2}$$

$$\left[\frac{1}{\sqrt{\{1 + [1/(r_3/r_1)]\}}} + \frac{1 - [1/(r_3/r_1)]}{2\,\{1 + [1/(r_3/r_1)]\}^{3/2}} - \frac{1}{2\sqrt{[2/(r_3/r_1)]}}\right]. \quad (5)$$

It is seen from Fig. 5 that the slope considered is positive for all values of the terminal ratio greater than or equal to one but less than the root of the above expression which is approximately equal to 15·58. This root is designated by the symbol ρ_{crit} (critical terminal ratio) and it may be shown by setting equation (5) equal to zero that ρ_{crit} is the positive root of the polynomial

$$x^3 - 15x^2 - 9x - 1\ .$$

The closed solution for this is

$$\rho_{\mathrm{crit}} = 5 + 4\sqrt{7}\cos\left(\frac{1}{3}\arccos\frac{1}{2}\sqrt{\frac{1369}{343}}\right), \quad (6)$$

and a more accurate evaluation yields

$$\rho_{\mathrm{crit}} \approx 15.5817187387632.$$

This may be shown to be exactly the same terminal ratio for which the Hohmann transfer modulus is a maximum. For values of the terminal ratio greater than ρ_{crit} the slope is negative, approaching zero from the negative side as the terminal ratio becomes infinite. Hence, it may be stated that the bi-elliptical transfer is more economical than the Hohmann transfer for some conjunction ratios at all values of the terminal ratio greater than ρ_{crit}.

Further investigation into the derivative of equation (4) will result in a more precise determination of the conjunction ratios for which the transfer modulus of exterior conjunction is smaller than the corresponding Hohmann transfer modulus. If the derivative of equation (4) with respect to the conjunction ratio is set equal to zero, the solution for the conjunction ratio as a function of the terminal ratio is

$$r_2/r_1 = -\frac{3 + [1/(r_3/r_1)]}{3\,\{1 + [1/(r_3/r_1)]\} - 2\sqrt{\{3 - [2/(r_3/r_1)]\}}}\ . \quad (7)$$

Physically, this solution is only applicable for exterior conjunctions—because of the assumptions in deriving equation (4)—and it may be shown that under this restriction, the right side of equation (7) may only assume values of the terminal ratio in the region

$$9 \leqslant r_3/r_1 \leqslant \rho_{\mathrm{crit}}\ . \quad (8)$$

Zero derivatives also occur at infinite conjunction ratios for *all* terminal ratios.

From this result, combined with the results concerning the slope of each of the parametric curves at their initial point, the following may be deduced:

(1) Because all parametric curves of terminal ratio greater than ρ_{crit} initiate with a *negative* slope and simultaneously possess *no* zero derivative for finite abcissa (conjunction ratio), the ordinate of these curves (transfer modulus) must be everywhere less than the ordinate of the initial point (Hohmann transfer modulus). In other words, *all bi-elliptical transfers of exterior conjunction are more economical than the Hohmann transfer for terminal ratios greater than* ρ_{crit}. Furthermore, since the slopes of these curves are continuously negative until the conjunction ratio becomes infinite, it follows that the profits in using the bi-elliptical transfer in this situation increase with the conjunction ratio, with the limiting profit at infinite conjunction ratio.

(2) It is obvious that the statements of (1) may be extended to include the terminal ratio ρ_{crit} itself.

(3) Parametric curves of terminal ratios in the closed interval from nine to ρ_{crit} possess one maximum and then decrease as the conjunction ratio becomes infinite. This is because the initial slopes of these curves are positive and they all possess a single zero derivative for finite conjunction ratios. The important point of interest

for these curves is whether any particular curve, after going through its maximum value, decreases to the point that the ordinate of the curve goes below the ordinate of the initial point. This will be discussed shortly.

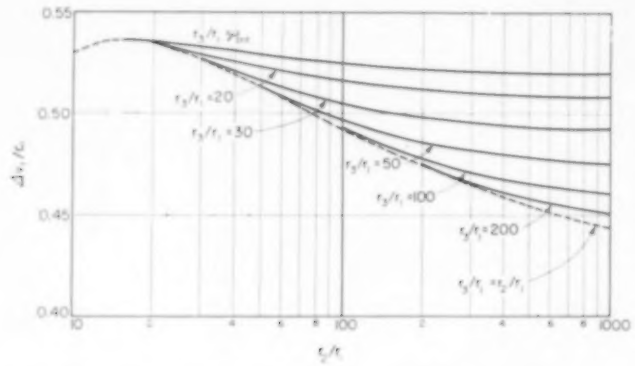


Fig. 6. Total velocity requirement (dimensionless) for bi-elliptical transfer if intermediate point is larger than terminal orbits for case that $r_3/r_1 > \rho_{crit}$.

(4) For terminal ratios less than nine (but greater than or equal to one) the initial slope is positive and continues to be so until the conjunction ratio becomes infinite, since, as has been seen, these curves possess no zero derivatives for finite conjunction ratios. Therefore, *all bi-ellipticals of exterior conjunction are less economical than the Hohmann transfer for terminal ratios less than nine.*

(5) The statement of (4) may be extended to include the terminal ratio of nine.

The curves of terminal ratio greater than or equal to ρ_{crit} are shown in Fig. 6, and the solution point for occurrence of zero derivatives in Fig. 7.

For curves of terminal ratios greater than nine and less than ρ_{crit}, it must now be determined which of these curves (if any), after going through their maximum values, decrease to values less than the magnitude of the initial point of the curve. The value of the ordinate of the first point of one of these curves is the transfer modulus for the Hohmann transfer and is given by equation (1) wherein the quantity r_2/r_1 must be replaced by the quantity r_3/r_1 because of the subsequent notational change. The value of the ordinate at any point on the parametric curve is of course given by equation (4).

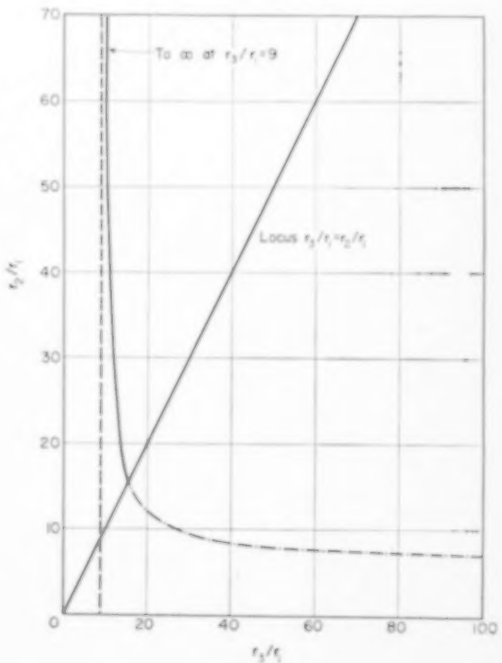


Fig. 7. Location of r_2/r_1 for which the total velocity increment ($\Delta v_t/c_1$) is a maximum plotted over r_3/r_1. — · — · — Solutions for which $r_3/r_1 > r_2/r_1$, which are therefore not applicable.

Any one of the parametric curves which possesses the property that it, at some point after reaching its maximum, becomes less than the Hohmann transfer modulus must yield a solution for the equality of equations (1) and (4) at some point other than the initial point, where the moduli expressed by equations (1) and (4) will always be equal. If equations (1) and (4) are set equal, the resulting equation is

$$\frac{1 - x_2}{\sqrt{[x_2(1 + x_2)]}} - \frac{1 - x_3}{\sqrt{[x_3(1 + x_3)]}} - \sqrt{\left(\frac{1}{x_2} + \frac{1}{x_3}\right)} + \sqrt{\frac{2}{x_3}} = 0 , \quad (9)$$

where $x_2 = r_2/r_1$, $x_3 = r_3/r_1$.

This possesses the solution

$$x_2 = r_2/r_1 = \frac{A + \sqrt{B}}{2} , \quad (10)$$

wherein

$$B = A^2 - \frac{4}{x_3[(M + R)^2 - 1]} ,$$

$$A = \frac{1}{x_3}\left[\frac{(M - 3R)^2 + 1}{(M + R)^2 - 1} + \frac{1}{x_3}\left(\frac{1}{1 - (M + R)^2}\right)\right],$$

$$M = R\left(K^2 - \frac{1}{x_3}\right), R = \frac{1}{2K} ,$$

$$K = \frac{x_3 - 1}{x_3}\sqrt{\left(\frac{x_3}{1 + x_3}\right)} + \sqrt{\frac{2}{x_3}} .$$

The other obvious solution to equation (9) is $x_2 = x_3$.

Equation (10) is shown in Fig. 8. The solution possesses the following properties:

(1) The conjunction ratio is infinite for a terminal ratio of ρ_f (feasibility ratio), where ρ_f may be defined as the largest positive root of the polynomial

$$x^3 - (7 + 4\sqrt{2})x^2 + (3 + 4\sqrt{2})x - 1$$

which has the closed solution

$$\rho_f = \frac{7 + 4\sqrt{2}}{3} + \frac{4}{3}\sqrt{(18 + 11\sqrt{2})}\cos\left[\frac{1}{3}\arccos\frac{1}{8}\sqrt{\left(\frac{4600970 - 162621\sqrt{2}}{68921}\right)}\right] .$$

A more precise determination yields

$$\rho_f \approx 11.9387654726 .$$

(2) The conjunction ratio is equal to ρ_{crit} when the terminal ratio is equal to ρ_{crit} and for values of the terminal ratio greater than ρ_{crit} the resulting conjunction ratio does not fall within the region of exterior conjunctions.

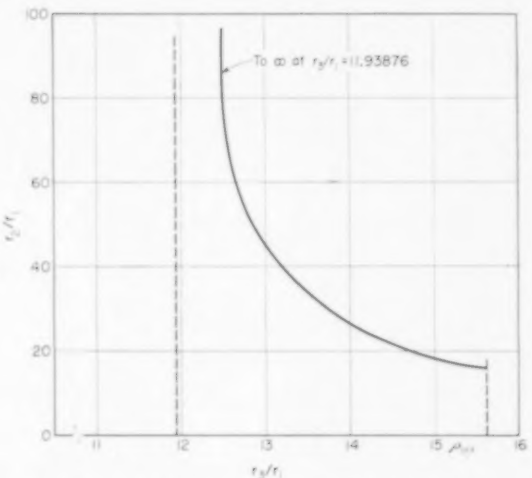


Fig. 8. Locus of feasibility for group $2{:}9 < r_3/r_1 < \rho_{crit}$.

(3) No real values of the conjunction can be obtained from equation (10) for terminal ratios less than ρ_f.

From these results the following statements may be verified:

(1) For parametric curves of terminal ratio greater than nine but less than or equal to ρ_f, the parametric curves, after going through their maximum values, decrease to some asymptotic minimum at infinite conjunction ratios, but this minimum is greater than the Hohmann transfer modulus except for a terminal ratio of ρ_f where it is equal to the Hohmann transfer modulus. Hence, all bi-elliptical transfers of exterior conjunction are less economical than Hohmann transfers for terminal ratios less than ρ_f but greater than nine. If this fact is combined with previous results, the following statement results.

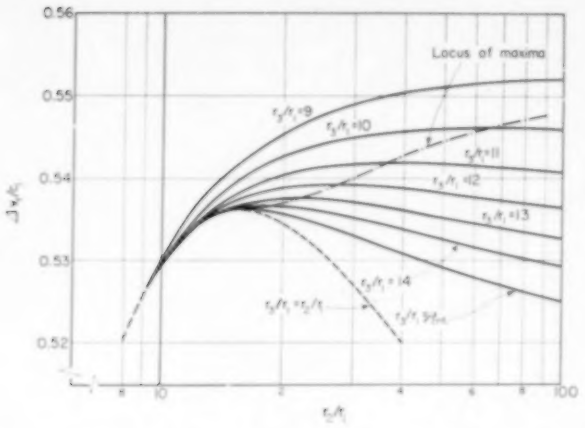


Fig. 9. Total velocity requirement (dimensionless) for bi-elliptical transfer, if intermediate point is larger than terminal orbits for case that $9 \leq r_3/r_1 \leq \rho_{crit}$.

All bi-elliptical transfers of exterior conjunction are less economical than the Hohmann transfer for terminal ratios less than ρ_f.

(2) *For terminal ratios greater than* ρ_f *but less than* ρ_{crit} *the bi-elliptical transfer is more economical than the Hohmann transfer provided the conjunction ratio is made greater than some critical value given by equation* (*10*) (Fig. 8). This critical value is infinite for terminal ratio of ρ_f and is equal to ρ_{crit} for a terminal ratio of ρ_{crit}. As in the preceding situations, maximum profits are obtained with infinite conjunction ratios. For this reason, the maximum profits for the bi-elliptical transfer are given by the differences in the ordinates of the curves in Fig. 2 showing the transfer modulus for an exterior conjunction of infinite radius and for the Hohmann transfer. This difference is shown by Fig. 10. Fig. 9 shows the parametric curves in the group with terminal ratios between ρ_f and ρ_{crit}.

The preceding discussions have covered all possible cases of exterior conjunction and have defined all situations for which bi-elliptical

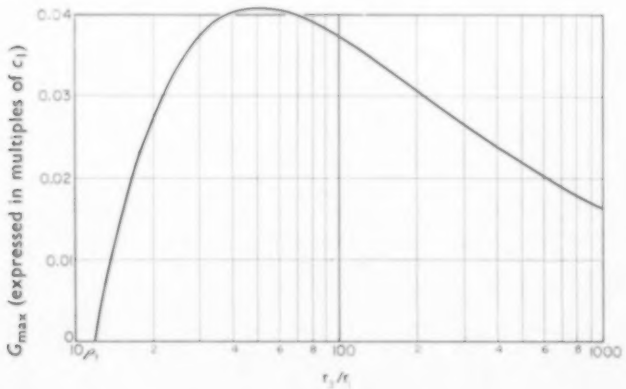


Fig. 10. Maximum possible gain for any given r_3/r_1 plotted over r_3/r_1.

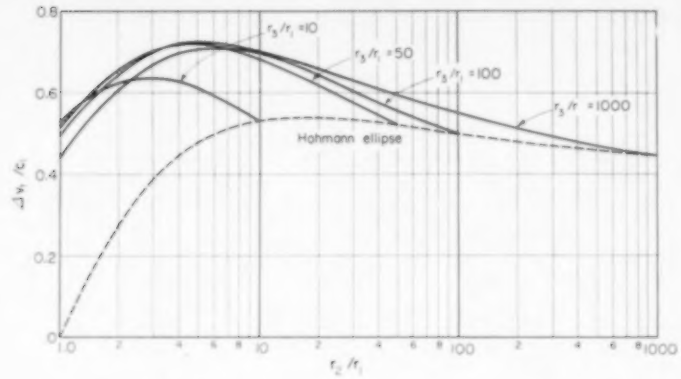


Fig. 11. Sum of velocity increments (dimensionless) for case of altitude of intermediate point between altitudes of terminal orbits $r_1 < r_2 < r_3$.

transfers are (in one sense) more favorable than the Hohmann transfer. It remains only to investigate interior conjunctions.

There exist two types of interior conjunctions, one for which the conjunction ratio is less than the terminal ratio but greater than one, and one for which the conjunction ratio is less than one. Transfer moduli for these two cases are shown in Figs. 11 and 12 respectively, and may be analyzed with very little effort.

For interior conjunctions with a conjunction ratio less than the terminal ratio but greater than

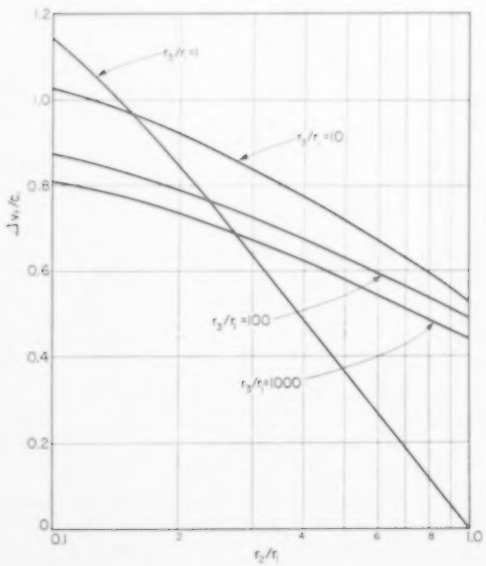


Fig. 12. Total dimensionless velocity increment vs. r_2/r_1 for various values of r_3/r_1; for case $r_2 < r_1 < r_3$.

unity, there exist the two boundaries where the conjunction ratio is equal to the terminal ratio or to unity. Both of these cases represent Hohmann transfer, and for this reason the two terminal points of the parametric curves in Fig. 11 represent Hohmann transfer moduli. Equation (3) may be re-written for this case without the absolute value signs if the positive orientations of the quantities in the absolute value signs are determined. If this expression is differentiated with respect to the conjunction ratio and evaluated at the point where the conjunction ratio is one, the derivative can be shown to remain positive for all terminal ratios greater than one, and zero at the terminal ratio of one. Furthermore, it may be shown that these curves possess *at most* one zero derivative which must therefore correspond to a maximum. Hence the transfer modulus along all of these curves must be greater than or equal to the Hohmann transfer modulus.

For the remaining case of interior conjunction, it may be shown that all of the curves in Fig. 12 initiate with a negative slope at the point where the conjunction ratio is equal to unity. Because the conjunction ratio for this case *decreases* to zero (neglecting the diameter of the attracting body), the transfer moduli begin to increase with variation of the conjunction ratio through its considered region. Also, it may be shown that the curves can possess no zero slope in this region so that they continue to increase in the value of the ordinate.

Fig. 13 shows one concrete application of the bi-elliptical transfer. For this case the terminal ratio was selected as ρ_{crit}.

CONCLUDING REMARKS

It is thought that the results herein present some light on the problem of optimum transfer between circular orbits, especially as regards the optimum number of impulses. Since minimum transfer moduli for the bi-elliptical transfer occur at infinity, it would seem that additional impulses would offer no real advantage. The

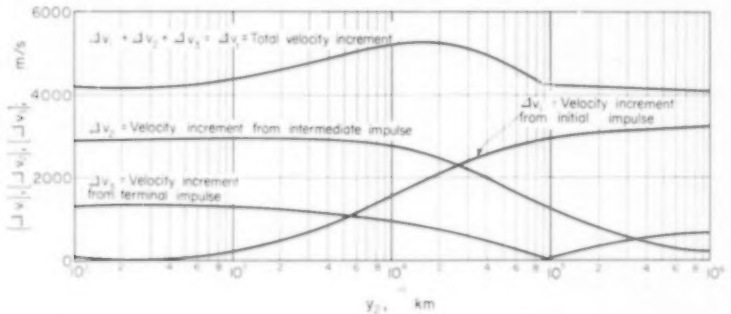


Fig. 13. Total velocity increments (Δv_t) and component velocity increments necessary for a bi-elliptical transfer (Δv_t) from 200 km circular orbit altitude to a 96,000 km circular orbit altitude vs. altitude of the intermediate point (y_2).

advantage of the case of three impulses over two has been seen to lie in the ability to obtain a very high-altitude intermediate point, which probably could not be accomplished any more economically by the application of more than three impulses.

Also, there are implications regarding transfers between non-co-planar orbits, as it would seem that for some situations it would be more economical to first obtain an intermediate point far out, make necessary inclination changes and then return to the desired orbit. This has been in part substantiated by Rider.[3]

Also, the entire investigation is increased by an order of magnitude when considering transfers between elliptical co-planar orbits. Preliminary analysis has confirmed that bi-elliptical

transfers are applicable and more economical than two-impulse transfers for some cases of transfer between co-planar elliptical orbits, and non co-planar as well.

Nomenclature

c_1 circular velocity of circular orbit of radius r_1
Δv velocity increment
Δv_t total velocity increment
r distance from point of interest to center of attracting body
ρ_{crit} positive root of $x^3 - 15x^2 - 9x - 1$
ρ_f largest positive root of $x^3 - (7 + 4\sqrt{2})x^2 + (3 + 4\sqrt{2})x - 1$

REFERENCES

1. W. HOHMANN, *Die Erreichbarkeit der Himmelskörper*, Oldenbourg, Munich (1925).
2. D. LAWDEN, *J. Brit. Interplanet. Soc.* **11**, 321 (1952).
3. L. RIDER, *Jet Propulsion* **29**, 48 (1959).

APPLICATIONS OF DYNAMIC PROGRAMMING TO SPACE GUIDANCE, SATELLITES AND TRAJECTORIES

RICHARD BELLMAN, STUART DREYFUS and ROBERT KALABA
The RAND Corporation, Santa Monica, California

1. INTRODUCTION

The feasibility of space travel and man-made satellites has created considerable interest in the determination of optimal trajectories and in guidance and control processes. These problems, long of purely mathematical and astronomical concern, have now become part of the engineering domain. The result is that there is a great demand for feasible numerical solutions of the associated analytic problems.

Many of these are classically of great difficulty. As a result of the intensive study of these questions, it is now well appreciated that the classical techniques of the calculus of variations are inoperative unless the problem is rather carefully selected.

In this paper we wish to sketch the applicability of a new mathematical technique, based on the theory of dynamic programming,[1, 2] to the computational solution of trajectory problems. Many problems, seemingly inaccessible to the conventional methods of the calculus of variations, have already been resolved.[2, 4–6, 10, 14]

Naturally, there are still a number of major difficulties to be overcome before it can be asserted that we possess a routine approach to realistic processes. What is important is that we have a general method for attacking guidance and control processes which permits us to study more complex physical problems using quite elementary mathematical concepts.

2. A PROPERTY OF OPTIMAL TIME TRAJECTORIES

Consider the problem of going from a point p in phase space to a point q in phase space. Symbolically, let this be represented in Fig. 1, where r is an intermediate point.

Let some particular trajectory be pursued, and denote by $t\,(p, q)$ the time required to traverse the path between two generic points. Then, clearly

$$t\,(p, q) = t\,(p, r) + t\,(r, q)\,. \qquad (2{\cdot}1)$$

Consider now the problem of determining, subject to various constraints on the allowable motions, the path of minimum time between p and q. Then, again clearly, if r is a point on an optimal trajectory, the times between p and r and r and q must also be minimum times for the respective points.

Fig. 1.

The problem we face is that of starting out from p. If we do not know the optimal trajectory *a priori*, the foregoing bit of information concerning geodesics is not particularly enlightening.

We can, however, make good use of the following part of this information. Wherever r is in phase space and whatever the time consumed going from p to r, the time involved in going from r to q must be a minimum for this part of the path.

This statement is a particular case of the "principle of optimality".[1]

3. DYNAMIC PROGRAMMING APPROACH [3, 4]

It remains to express the foregoing ideas in analytic form. Consider

$$f(p) = \text{the time consumed in going from an arbitrary point } p \text{ to a fixed point } q, \text{ using an optimal path.} \tag{3.1}$$

Then, the principle of optimality yields firstly the equation

$$f(p) \leqq t(p, r) + f(r), \tag{3.2}$$

for any path $\overline{prq}$. To determine r, we minimize over all admissible choices of r, obtaining the basic equation

$$f(p) = \min_{r} [t(p, r) + f(r)]. \tag{3.3}$$

This equation is the basis for the computational solution of trajectory problems.

Similar considerations lead to the determination of kth best paths, as is discussed in Ref (7).

4. BRACHISTOCHRONE

As an example, let us consider the classical brachistochrone problem. Given two points, P and Q, in a vertical plane, we wish to determine a curve connecting these two points with the property that a particle sliding down this curve, solely under the influence of gravity and with no frictional forces, will traverse the path in minimum time.

Take P to the origin and let Q be the fixed point (x_0, y_0), as indicated below.

Omitting the gravitational constant, of no import, the problem in the calculus of variations is that of minimizing the integral

$$J(y) = \int_0^{x_0} \left(\frac{1 + y'^2}{y}\right)^{1/2} dx \tag{4.1}$$

over all curves satisfying the end conditions $y(0) = 0$, $y(x_0) = y_0$. In place of treating this in the usual fashion, *à la* Euler equation and so on, let us introduce the function $f(x, y)$ defined as follows:

$$f(x, y) = \text{the time required for a particle to fall from } (x, y) \text{ to } (x_0, y_0) \text{ along an optimal path.} \tag{4.2}$$

Then a particularization of equation (3.3) yields the functional equation

$$f(x, y) = \min_{y'} \left[\left(\frac{1 + y'^2}{y}\right)^{1/2} \Delta + f(x + \Delta, y + y'\Delta) + 0(\Delta) \right] \tag{4.3}$$

where y' represents the slope of a proposed path at x, and Δ is an infinitesimal.

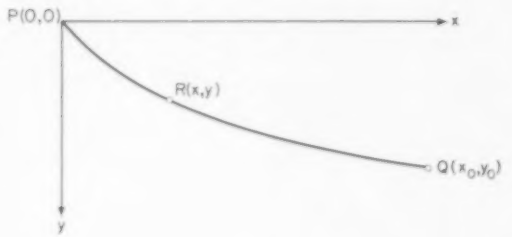


Fig. 2.

Passing to the limit, we can obtain a partial differential equation for f which turns out to be equivalent to the Euler equation. But this is not what we wish to do! Instead we sub-divide the x-interval into N parts, where $N\Delta = x$ and regard $f(k\Delta, y) = f_k(y)$ as a function of y for $0 \leqq y \leqq y_0$.

Then equation (4.3) becomes

$$f_k(y) = \min_{y'} \left[\left(\frac{1 + y'^2}{y}\right)^{1/2} \Delta + f_{k+1}(y + y'\Delta) \right], \tag{4.4}$$

for $k = 0, 1, 2, \ldots, N - 1$, with $f_N(y_0) = 0$. The determination of $f_k(y)$ is a computational process which we refer to a digital computer. The minimization over initial directions is carried out not by means of calculus, but by means of a direct search process over allowable slopes.

One consequence of this is that constraints on y' simplify the computational process, as they should since we have fewer policies to examine. Further details of the computational aspects, which are seldom trivial, will be found in Ref. (10).

5. MINIMUM ENERGY

Suppose that we wish to go from p to q in phase space, reverting to the general formulation of Section 3, at minimum cost in resources, say minimum fuel. Regarding $t(p, r)$ not as a time, but as a cost for getting from p to r, we see that the general equation of (3.3) is still valid.

Generally, we can treat any variational problem in this fashion, providing it has the Markovian property that we have been employing. To use the foregoing method we need merely know that the optimal continuation from any point in phase space depends only upon that point, and not upon the past history of the process.

Thus, not only can general variational problems involving standard functionals of the form

$$J(y) = \int_{x_0}^{x_1} g(x, y, y')\,dx, \qquad (5.1)$$

be handled but also many classes of implicit variational problems[5, 9] which cannot readily be treated by conventional methods.

6. STOCHASTIC AND ADAPTIVE CONTROL PROCESSES

Finally, the same general techniques can be used to treat the more complex stochastic and adaptive control processes.[8, 9, 11, 12]

7. DISCUSSION

The trajectory and feedback control problems sketched in earlier sections represent merely one application, albeit a particularly important one, of the functional equation technique of dynamic programming to the solution of space problems. Problems which lead to multi-stage decision processes also occur in the areas of design of multi-stage rockets,[14] communication,[3, 15] propagation,[13] equipment reliability[16] equipment replacement,[17, 18] and so on.

The desire for the design of more nearly optimal systems coupled with the inevitable advances in computer technology should bring dynamic programming methodology into a position of increasing importance in the future in the resolution of space optimization problems.

REFERENCES

1. R. Bellman, *Dynamic Programming*, Princeton Univ. Press (1957).
2. T. Cartaino and S. Dreyfus, Application of dynamic programming to the airplane minimum time-to-climb problem, *Aeronaut. Rev.* (1957).
3. R. Bellman, *Quart. Appl. Math.* **16,** 87 (1958).
4. R. Kalaba, *On Some Communication Network Problems*, The RAND Corp., Paper P-1325 (Apr. 1958).
5. R. Bellman and J. M. Richardson, *Quart. Appl. Math.* **17,** 231 (1959).
6. R. Bellman and S. Dreyfus, *J. Brit. Interplan. Soc.* **17,** 78 (1959–60).
7. R. Bellman and R. Kalaba, *On k-th Best Policies, J. Soc. Ind. Appl. Math.*, in press.
8. R. Bellman and R. Kalaba, *Proc. Nat. Acad. Sci., Wash.* **45,** 1288 (1959).
9. R. Bellman, *Adaptive Control. A Guided Tour*, Princeton Univ. Press, to be published.
10. R. Bellman and S. Dreyfus, *Computational Aspects of Dynamic Programming*, to be published.
11. R. Bellman and R. Kalaba, *Trans. Inst. Radio Engrs. on Autom. Control*, vol. AC–5, pp. 5–10 (1960).
12. R. Bellman, Dynamic programming and stochastic control processes in *Information and Control*, Vol 1, pp. 228–239 (1958).
13. R. Bellman and R. Kalaba, *Trans. Inst. Radio Engrs. on Circuit Theory* **6,** Spec. Suppl. p. 271 (May 1959).
14. R. P. Ten Dyke, *Jet Propulsion* **28,** 338 (1958).
15. R. Bellman and R. Kalaba, *Trans. Inst. Radio Engrs. on Inform. Theory* **3,** 197 (1957).
16. R. Bellman and S. Dreyfus, *Operat. Res.* **6,** 200 (1958).
17. R. Bellman, *J. Soc. Ind. Appl. Math.* **3,** 133 (1955).
18. R. Kalaba and M. Juncosa, General Systems Approaches to Telecommunication Optimization Problems, *Inst. Radio Engrs. Nat. Convent. Rec.*, Part 8, p. 203 (1957).

MISSILE SIMULATION ON AN ELECTRONIC DIGITAL COMPUTER

A. HASKELL EPSTEIN and NORMAN W. PETERSON
Computation and Data Reduction Center, Space Technology Laboratories, Inc.

Abstract—A large-scale digital computer can be a powerful and useful tool to a systems engineer for evaluating and studying proposed guided missile systems. As mathematical models for each of the subsystems are developed, they can be programmed and checked independently. When operating, they can be combined into a computer program capable of simulating the complete missile system under the expected flight conditions. Several applications of missile simulation programs are presented along with a generalized description of the methods that are employed to produce computer programs at the Space Technology Laboratories.

Since antiquity, models have been used in studying physical and mathematical phenomena. The models familiar to us from geometry immediately come to mind. Models of the plain or solid geometry sort are usually passive in nature. Dynamic models also have been used for centuries. Planetary models for predicting tides and astrological configurations have been in use at least since Grecian times.

Modern engineering models are no less in evidence. Wind tunnels and model basins are vital parts of the aircraft and shipbuilding industries. For ballistic missiles, wind tunnels and model basins also have their part. But a very large area of function of ballistic missiles cannot be modeled on any smaller physical scale than the missile system or subsystem itself.

In order to make effective use of passive or dynamic models, a precise understanding of the mathematical description of the object system is necessary. Usually the object system must be mathematically described prior to modeling. The mathematics must then be reduced to a physical state, perhaps in an entirely different realm than the original system. Thus, a mathematical system may have its model (or analog) in an electrical system. Indeed, many large areas of research would be crippled without electronic analog computers.

Another area of modeling has appeared in recent years with the advent of large-scale digital computers. These computers are quite different in philosophy from the previously mentioned analog computers in that they circumvent the final reversion to the physical state. Once the mathematical definition of the object system is known, no further resort to models need be taken.

A digital computer in itself can perform only arithmetic operations and logical operations; that is, it can usually add, subtract, multiply, divide and change operation sequences upon finding that one number is larger than another number. These computers perform their operations upon internally stored numbers with an internally stored sequence of instructions. This sequence of instructions may be quite large (tens of thousands) and will be self-modifying. The precision with which a digital computer performs its arithmetic computations is limited only by the size of its internal registers which may be effectively varied by programming techniques. This is not true in the case of analog computers which depend on the precision with which measurements of physical quantities may be taken.

The sequences of instructions for a digital computer are called programs. Programs may interpret missile subsystems in computer language; that is, a subsystem of a missile may be simulated (modeled) by a certain group of computer instructions. These groups of computer instructions are pyramided to form a complete simulation of a missile—much as the

subsystems themselves are pyramided to form a missile system.

Simulations of ballistic missile systems and their environments are particularly useful in that most older methods of design and design check fail when applied to a system of the complexity of an ICBM. "Cut and try" methods are obviously not feasible. They are too expensive at ten million dollars per shot. Analytic methods of design are, of course, desirable. However, analytic methods usually depend on a high degree of system linearity and simplicity. Many things about ballistic missiles are nonlinear. Also, because of their complexity, ballistic missile parameters are difficult to handle analytically—just the bookkeeping would become quite difficult.

Despite these and other difficulties, methods must be found for specifying the constraints on a subsystem, the necessary output from that subsystem and the environment in which the subsystem must operate. One must be able to check the influences of one known subsystem on a proposed subsystem. Indeed, one must be able to predict the interactions of all subsystems. Digital simulations of subsystems and systems of subsystems can, to a considerable extent, provide these answers.

When the need for a subsystem is appreciated, the input–output characteristics (transfer function) required for that subsystem may be precisely delineated by iterative methods. In iterating to find these characteristics, full advantage can be taken of known characteristics, linear or nonlinear, of all other subsystems. The only constraints are that the desired total performance be precisely known and stated in mathematical language.

Entire trajectories may be flown on a digital computer to determine the effect of perturbations of any sort that can be precisely stated. For instance, the effects of winds, gyro drift, radar noise, thrust decay, thrust build-up and slight engine malfunctions may be included in one perturbed trajectory. If it is desired to find an initial pitching rate which will produce as long a trajectory as possible, or one with miss coefficients at burnout as low as possible or with certain mass heating requirements, the answer may usually be found by iterative methods. One usually needs to provide a suggestion to the computer program as to what to do in the case of missing the desired characteristic in a positive or a negative fashion. If the suggestion is correct and not too many parameters are being optimized at once, the iteration will usually converge to a useful result. Speed of convergence may be increased by more sophisticated and powerful techniques than simple random or suggested perturbations.

Detailed simulations are used to produce trajectory reports for agencies interested in a specific missile test. In simulations intended for this purpose, a large amount of output relating to subsystems of various sorts is included. Frequently, agencies involved with subsystems which are not directly simulated get useful information. If their subsystems do not affect any of the simulated subsystems, lack of inclusion will not be detrimental to the total of the simulation.

One of the useful facets of digital computers is their ability to simulate other digital computers. Since digital computers are being used with increasing frequency in missile guidance systems, this facet leads to simulations including: the missile, a radar system and a guidance digital computer. Digital computers can also simulate analog computers and, therefore, analog guidance systems or analog control systems.

Since digital computers, unlike their analog cousins, cannot directly perform the calculus operations, integration must be performed by numerical methods. In one of these methods, derivatives are needed at the point of departure and at certain points between the point of departure and a later time increment. Thus, a digitally simulated missile trajectory, rather than being a smooth, continuous thing as it is in real life, is rather a series of discrete points. The length of time between a pair of these points is known as the step size. The upper bound on the magnitude of the step size, in general, depends on the rapidity with which the derivatives are changing. The lower bound is limited by round-off error and computer time considerations.

The upper bound is associated with the "truncated series" nature of the integration method. Round-off error is due partly to the loss of significance when adding numbers of different sign and magnitude. Ordinarily, step sizes for missile trajectories are in the low-order number of seconds. Step sizes much smaller than this, perhaps as low as a hundredth of a second, are sometimes necessary when subsystems of comparatively high frequency, such as autopilots, are involved.

The discrete nature of numerical integration makes this method fit well with digitally guided missiles since the guidance computers also have this "discrete" time characteristic. When analog computers are simulated, it is a matter of performing the analog computer operations with the subsystem simulations which compute the derivatives used by the integration procedure.

Physically, integration is used primarily to solve Newton's second law for velocity and position. That is, simulations are predicated on the assumption that, from any known velocity and position of a missile, the forces can be computed which act at that point and which will determine the position and velocity of the missile a short time later.

Several engineering groups frequently use the same digital simulation. Although many of those groups will have co-ordinate systems peculiar to their own groups, the missile position and velocity only, need be integrated in a co-ordinate system convenient to the simulation. All other co-ordinate systems may be derived from this co-ordinate system. A co-ordinate system in which the rotation of the Earth is depicted (with its movement of the launch pad, target and radar system) is usually necessary. From this co-ordinate system all others may be obtained: a Cartesian system centered at the launch pad, a spherical system centered at the radar antenna, latitude and longitude of the missile. All these might be necessary to specify the position of the missile to the various groups interested. The velocity of the missile, likewise, may be specified in numerous ways. Its total velocity with reference to the center of the Earth at the time of launch may be interesting. Of further interest is the magnitude of the velocity with respect to air. In order to compute the aerodynamic drag, the longitudinal component of the latter must be known, also the normal velocity of air may be needed in a more sophisticated simulation. The velocity of the missile with respect to the radar antenna may also be necessary.

One system which provides all of this information implicitly is the Earth-centered inertial system (ECI). In this system, the center of the Earth is the center of a Cartesian three-dimensional system in which the Earth rotates about one axis. Once the position of the missile on its launch pad is converted from, for instance, latitude and longitude to ECI co-ordinates, no further reference to geodetic quantities need be made (except, perhaps, for gravity).

An extension of this system applies to the attitude of the missile. If a three-dimensional vector system is applied at the missile, one vector may lie in the axis of the missile, a second vector may lie in the pitch plane, and a third vector may lie in the yaw plane. They are all mutually perpendicular to each other. These vectors may be unitized, in which case they are useful in resolving various force vectors into their Cartesian components. If steering is to be simulated, the vectors representing the attitude of the missile become variables of integration.

When this universal co-ordinate system is in effect, all program building blocks, or subsystems, will relate to each other through this system. Thus, pools of these program subsystems may be accumulated and may be drawn upon in making any simulations or in adding any subsystems to old simulations.

In making large-scale simulations, it is frequently found that whole missile systems may be regarded as subsystems. This may be useful for iterative purposes in investigating trajectories or perhaps for missile combination studies.

In forming the actual subsystem simulation, the first problem is to state the nature of the subsystem—with numerical or analytic method—so that the computer can be coded to resemble the subsystem. This may take the form of the use of a simple empirical table. For instance, thrust and propellant flow may be simulated as

a table of thrust vs. atmospheric pressure or as a table of thrust vs. time, etc. Where more sophisticated equations are available, analytic methods for simulation may be applied. For instance, an analytic radar simulation is really nothing more than the co-ordinate transformation from ECI to spherical co-ordinates centered at the radar. Indeed, even such a thing as thrust may be simulated in a somewhat analytic method. Of course, the complete thermodynamic equations of a rocket engine might be programmed, but this would be a lengthy project. More practical program might well be one utilizing test stand data to correct a table of nominal thrust. Thus, influence coefficients may be found which relate thrust to air pressure, acceleration, etc.

When the mathematical model has been defined, the first step toward digital simulation has been accomplished. Digital computers in themselves are useful only when provided with a program. Programs are groups of machine instructions that perform a function. These machine instructions implement the arithmetic operations and also implement logical operations such as a transfer of logical control. While general-purpose digital computers are quite varied, they have one feature in common, that is, their ability to modify their own programs. This feature, together with conditionally controlled transfers, make possible a coding technique known as looping. In this technique, a sequence of instructions, sometimes quite short, is executed many times. Each time, slight changes are made in the portion of the instructions which refer to specific numbers. In this manner, constants (or variables) may be exchanged, while equations remain the same. This technique effectively allows one to execute millions of instructions in the course of a machine run while there are only thousands of instructions in the machine. This self-modifying ability found in digital computers is their principal source of computing power. To implement a sequence of algebraic equations, or formulas, may require several instructions or several dozen instructions depending upon the machine characteristics and upon whether the equations are abstractly similar enough that advantage may be taken of looping techniques.

In the machine, instructions appear as binary numbers. However, no reference need be made to the binary. Programs are available which perform transformations from English mnemonic operation code to binary operation codes. In some instances, programs may be used to translate mathematical language directly into sequences of operation codes. This, in effect, computers are able to code for themselves. The first form of coding aid is known as a symbolic assembler and the second is known as an algebraic compiler. Both of these two forms of coding aids are used in producing subsystem simulations. Compilers may also be used to combine subsystems into entire system simulations. This latter feature is particularly useful in that libraries of subsystem simulations may run into the several hundreds of current simulations.

The subsystem technique is used for far more than simulation of physical subsystems. Use is made of programs which resemble subsystems to perform operations peculiar to the simulation itself. For instance, the integration program itself resembles a subsystem, yet is not found in the missile. Indeed, large libraries of mathematical subsystems are generally available and may include routines that vary in power from complex eigenvalue programs to a simple sine or cosine subroutine. Part of the power of simulation is due to the accumulation of such mathematical tools. In the years to come tools of ever-increasing power can be expected.

Once a program is prepared, it may be run time after time on the machine with changes only in the input quantities. However, more typically, the algebra and logic also change, perhaps daily. These changes, if simple enough, may be accomplished without resort to the relatively expensive compiler or symbolic coding techniques, that is, the skilled programmer may insert these changes into the machine directly in binary. In some cases, the changes may be inserted symbolically but without the time consumption involved in the initial symbolic assembly.

One problem not directly connected with simulation which will plague a group using a simulation as a tool is that of record-keeping. It is frequently imperative to know exactly what went into a trajectory that was run months ago.

The actual output of the trajectory run may not be sufficient for this purpose. The machine can keep track of this information, storing it in an off-line storage medium such as punched cards or magnetic tape. This storage, however, creates library problems and, thus, library searching programs are also necessary. These records may be kept in a less automated manner by using off-line card equipment to produce tabular listings of input quantities and binary changes for every run.

If some record technique is not used, then complete input must be supplied for every run. With record techniques, one may combine past runs with new changes and thus minimize the need for voluminous communication between the programmer or person using the routine and the engineer or person utilizing the results.

The output formats of a digital simulation are, of course, limited by the output mechanisms associated with a given digital machine. While simple 120-column tabular listings are common, microfilmed graphs and typewritten sheets also are occasionally seen.

To the engineering group using a digital simulation then, the program appears only as mathematical equations and logic with associated input and output numbers. The accuracy of the simulation depends on the sophistication of the mathematical model and the register size of the machine.

The total power of digital simulation as a design and analysis tool cannot but increase as larger and faster digital machines permit larger and more sophisticated simulation programs, with the machine itself performing an ever-increasing share of the work involved in formulating a simulation.

BIAX HIGH SPEED MAGNETIC COMPUTER ELEMENT*

C. L. WANLASS and S. DEAN WANLASS
Aeronutronic, a Division of Ford Motor Company

Abstract—The BIAX digital computer elements are completely new magnetic memory and logic elements. They are multi-aperture devices fabricated out of conventional ferrite materials and are, generally speaking, an order of magnitude faster than present magnetic devices. To attain these speeds, they employ a fundamentally different principle of operation than is used with existing ferrite core and multi-aperture units. The BIAX principle employed is one of flux interference, not flux steering as is common in most multi-aperture devices. In addition to increased speeds, the BIAX elements also have important environmental advantages and are capable of operation at high temperatures. Another important advantage of the elements is their non-destructive read characteristic. Important cost savings also appear to be achievable with these units, particularly in terms of the logic elements.

These BIAX memory and logic elements are the result of approximately two and one half years of work conducted in Aeronutronic's Computer Research Laboratory. The elements have been reduced to practice and are currently in production. Aeronutronic is applying these BIAX techniques to a number of commercial and military projects where multi-megacycle data processing is a requirement.

BIAX MEMORY ELEMENT

Introduction

The basic BIAX memory element is shown in Fig. 1. It is a pressed ferrite element having two non-intersecting orthogonal holes. The basic concept employed is that of flux interference which occurs in the magnetic material located between the two orthogonal holes. The flux around each hole must share this common volume of ferrite material and a change in the flux linkage around one hole will cause a change in the flux linking the other hole. The degree of the flux interference can be controlled by varying the distance between the two orthogonal holes. In the basic BIAX memory element, the hole spacing is so chosen that this interference is reversible. Thus, a non-destructive memory element is achieved.

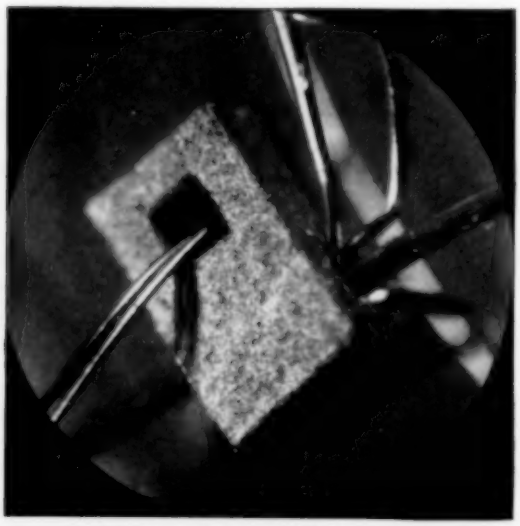


Fig. 1. Production BIAX element.

The element shown in Fig. 1 is a production BIAX element. Its physical dimensions are 50 × 50 × 85-mils with two 20-mil square holes. The distance between the two holes is the same as the distance between the inner and outer walls of each hole.

Theory of Operation

The theory of operation of the BIAX memory element can be explained with the aid of Fig. 2. For purposes of discussion, the upper hole will be referred to as the storage hole and the lower hole will be referred to as the interrogate hole. The storage hole contains the conductors associated with writing into the memory element and the sense winding. The interrogate hole contains the conductors associated with reading or interrogation of the element.

* Also presented at 1959 WESCON, San Francisco, Calif.

As has been stated, the distance between the two holes and the outer wall thicknesses are approximately the same. Under these conditions, complete magnetic saturation can be achieved around a given hole only if all the domains in the volume between the two holes are favorably oriented. Thus, if one attempts to saturate both holes simultaneously, a condition such as that associated with the interrogation currents were oriented in the same direction as the flux already established around the interrogate hole, then this interrogate current would cause domains in the common volume to be momentarily re-oriented in favor of augmenting the flux around the interrogate hole. This reorientation would, of course, decrease the flux linkages around the

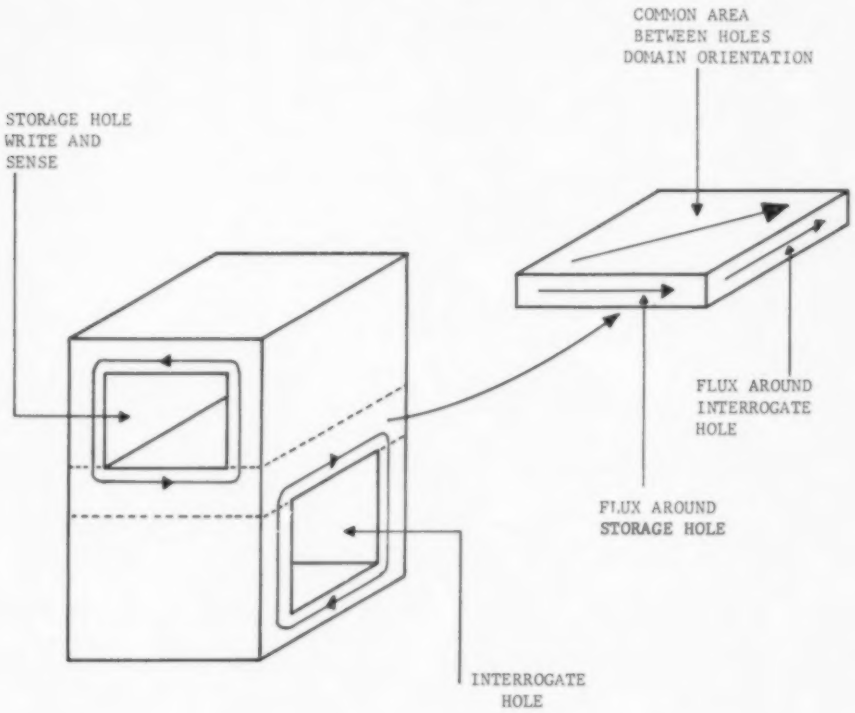


Fig. 2. BIAX memory element.

shown in Fig. 2 exists and the flux linking each hole is reduced to approximately $\frac{7}{10}$ of that which could be attained if one of the holes had no net flux linking it.

To explain the operation of the BIAX memory element more fully, assume that the common volume between the two holes is always saturated. Under these circumstances, it is possible to explain the memory operation in terms of the resultant domain orientation associated with this common volume (see Fig. 2). Assuming the general conditions shown in Fig. 2 prevailed at the time of interrogation and that the flux storage hole and this $\Delta\phi/\Delta t$ would give rise to a voltage on the sense winding passing through the storage hole. It is important to note that the sign of this output voltage will be solely dependent (for a given mode of interrogation) on the orientation of the flux encircling the upper hole at the time of the interrogation; thereby providing the basis for binary storage. After the interrogate current has terminated, the original conditions shown in Fig. 2 once more prevail, and a truly non-destructive interrogation has been achieved.

Obviously, other variations in terms of initial

conditions and the sign of the interrogation currents are possible and if these are examined in the light of the example cited, they will provide additional appreciation for the basic BIAX memory concept.

BIAX Write Characteristics

To store information in the BIAX memory element, current pulses of the proper sign, magnitude and duration pass through the storage hole so that the residual flux around the storage hole corresponds to the plus B_r or minus B_r point on the hysteresis loop (see Fig. 3). By definition, these conditions represent the storage of a "1" or a "0", respectively. Since the BIAX memory element has a non-destructive read characteristic, writing is only necessary when the stored information needs to be changed.

A sketch of the BIAX memory element along with its proper complement of read, write and sense conductors is shown in Fig. 4. The precise number of wires associated with the write hole is dependent upon the write system that is employed. For example, a coincident current write system could be used but this is not advisable because this system has definite speed limitations and is sensitive to material characteristics. In order to switch the BIAX element faster than is possible by conventional coincident current techniques, it is necessary to increase the magnitude of the writing current in order that the flux swing can occur in shorter intervals of time. A typical curve showing switching time as a function of the applied field (or write current) is shown in Fig. 5.

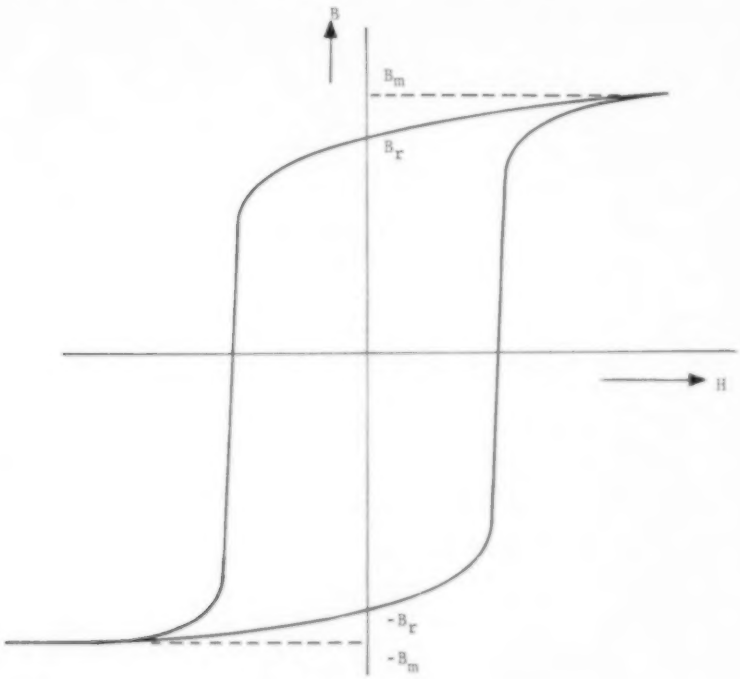


Fig. 3. Hysteresis loop.

The write system generally employed with BIAX memory arrays is referred to as the oersted second write system. This is a coincident write system but is basically different from the conventional coincident current system. This difference has been depicted on Fig. 5 and is discussed in more detail below.

Associated with high speed write techniques such as the oersted second write system, there

exists a basic write-disturb problem. This arises because any current greater than that corresponding to the low frequency coercive value will cause some flux change in the element. For example, with typical ferrite materials, one-half of the current required to switch a small ferrite element in 0·2 μsec can cause a partial reset from the initial residual flux condition. With conventional core memory array systems this would present a serious problem since during read-out a large noise output would occur and result in a poor signal-to-noise ratio. However, due to the bipolar output characteristic, BIAX elements have no basic signal-to-noise problem, and the only significant requirement for storage is that the residual flux condition does not change sign after an infinite number of disturb currents. The fulfillment of this condition has been demonstrated satisfactorily for the BIAX memory element described and thus, the only effect of a normal disturb current in the storage hole is to reduce the magnitude of the output signal generated during interrogation. This reduction in output depends on the magnitude

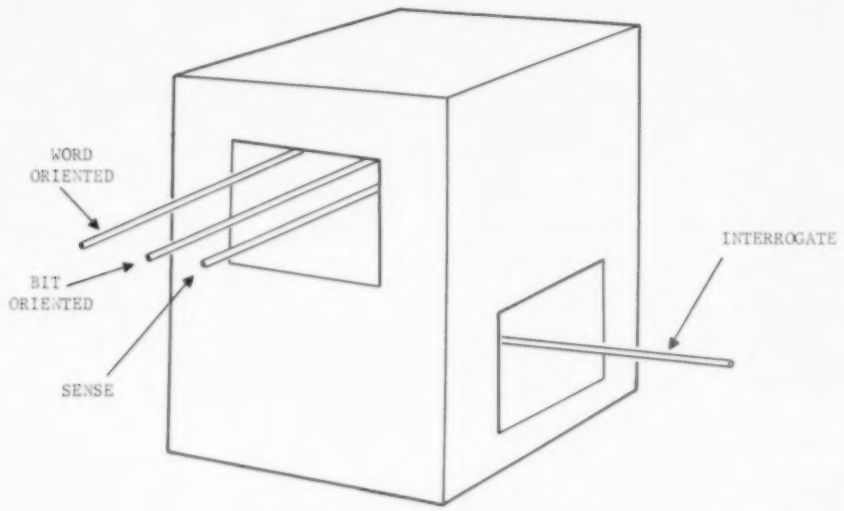


Fig. 4. Wired memory element.

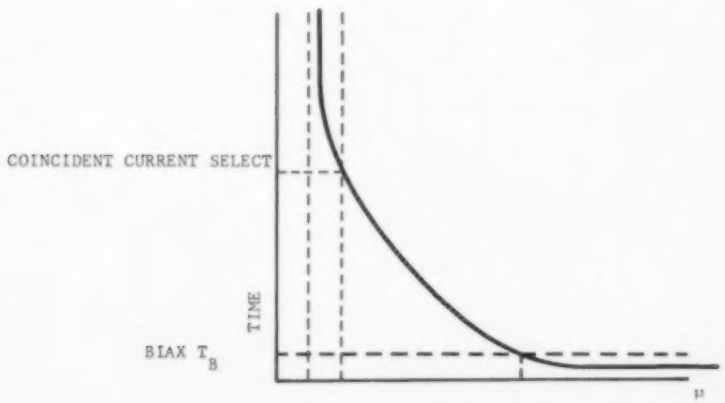


Fig. 5. Switching time vs. applied field.

and duration of the disturb current, the number of times the disturb occurs and the type of ferrite material employed.

The extent of the write disturb problem as applied to a standard BIAX element is shown graphically in Fig. 6 for typical cases. The curves presented in this figure are for a write sequence consisting of a single full write current followed by an infinite number of disturb pulses

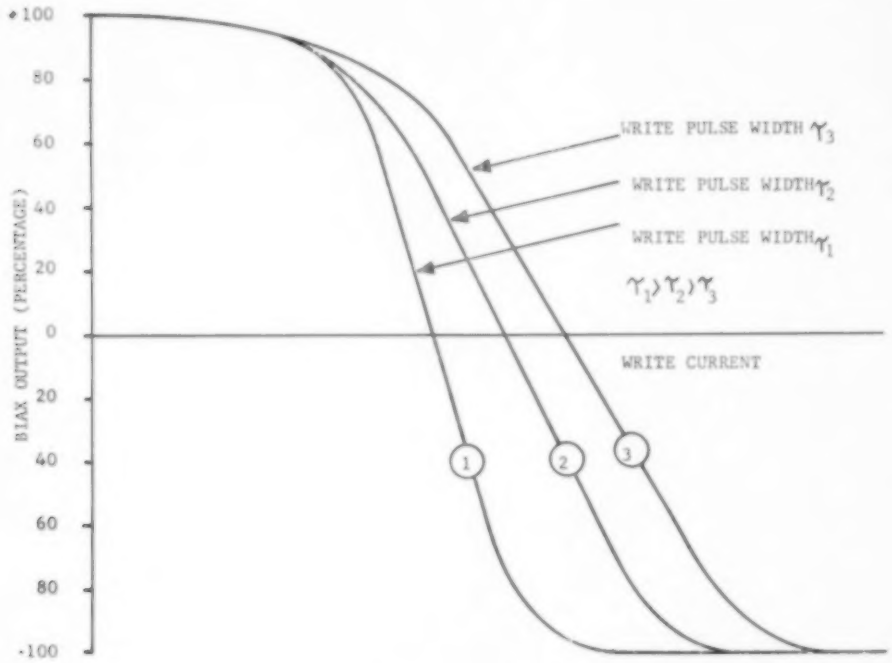


Fig. 6. BIAX write disturb characteristics.

of the magnitude indicated by the X-co-ordinate. (The latter having polarities opposite to the initial full write current pulse.) These curves also illustrate the bipolar BIAX output possible using conventional coincident current techniques. In this regard, curve 1 shows the write-disturb characteristics for a system employing a pulse width approximately equal to that of a coincident current type write system, while curves 2 and 3 show the corresponding disturb characteristics for write currents with considerably smaller pulse widths. (Curve 3 corresponds to pulse widths of approximately $\frac{2}{10}$ of the normal switching time.) These curves demonstrate that a reliable write system is possible at high speeds even when adequate consideration is given to such practical factors as drive current characteristics and variations in individual BIAX element. Using BIAX techniques, it is possible to employ a soft magnetic material and narrow pulse widths to achieve a condition corresponding to curve 3 of Fig. 6.

The photographs which constitute Fig. 7 show the high speed write and read capability of the BIAX element. The first trace of (d) shows the write current employed for a full write "1". The second trace of (d) shows the partial write "0" and full write "0" currents for this same system. The top trace of photo (e) shows the flux swings associated with the write hole. The bottom trace of (e) shows the currents associated with the interrogate axis. The top trace of photo (f) shows the BIAX "0–1–0" output. The large signal between interrogate current pulses is associated with the write

current. Directly above the center interrogate current is shown the "1" output. The signals above the other interrogate currents show the output for a "0" condition.

In this case, the system employed requires a positive signal to be a stored "1" and no signal, or a negative signal, to be a stored "0" output. This is done to allow an even greater safety factor for write-disturb currents with the use of soft magnetic materials. It should be noted that a full write "1" requires only 400 mA turns and a full write "0" approximately 240 mA turns. (The materials employed in subsequent systems require substantially less current due to the use of softer materials.) This system of write allows large tolerance on BIAX elements and drive currents with complete reliability. It should also be noted the write "1" read, write "0" read, write "1" read, sequence is occurring within a $4\frac{1}{2}$ μsec time interval.

The above data are sufficient to demonstrate that a high speed write system is feasible for a BIAX element employing conventional ferrite magnetic materials. Furthermore, it has also been demonstrated in the laboratory that reasonable variations in BIAX element characteristics and driver currents can be tolerated. In addition, it is important to note that transistorized drivers may be used at these frequencies because of the reasonable driver current requirements and low back voltage interrogate characteristics.

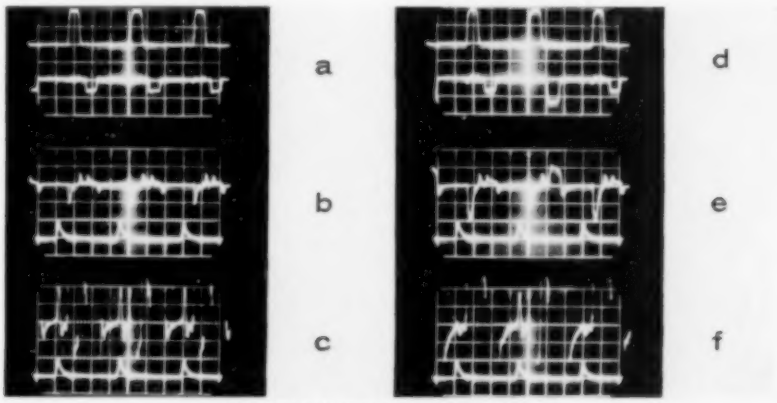


Fig. 7. High speed memory operation.
(a-upper) Write "1", 200 mA-turn/vert. div.
(a-lower) Partial write "0", 200 mA-turn/vert. div.
(b-upper) Flux swing in write hole, 0·4 V per turn/vert. div.
(b-lower) Interrogate currents, 200 mA-turn/vert. div.
(c-upper) BIAX "1" outputs, 10 mV per turn/vert. div.
(c-lower) Interrogate currents, 200 mA-turn/vert. div.
(d-upper) Write "1", 200 mA-turn/vert. div.
(d-lower) Partial write "0", write "0", 200 mA-turn/vert. div.
(e-upper) Flux swing in write hole, 0·4 V per turn/vert. div.
(e-lower) Interrogate current, 200 mA-turn/vert. div.
(f-upper) BIAX "0–1–0" outputs, 10 mV per turn/vert. div.
(f-lower) Interrogate current, 200 mA-turn/vert. div.
Horizontal scale for all traces: 0·4 μsec/div.

Later in the paper a BIAX memory system will be discussed which has proved feasible with respect to write disturb and non-destructive readout and which employs semiconductor drivers, sense amplifiers and decoders.

BIAX Read Characteristics

It is evident from the preceding discussion that the BIAX write system could be employed in conventional magnetic cores. However, it is only when these write techniques are coupled with the BIAX read techniques that a greatly

enhanced system is achieved. The BIAX read principle minimizes the write-disturb problem and also allows the use of minor hysteresis loop operation which is effective in reducing the volume of magnetic material employed in the read and write operation. Minor loop operation also greatly reduces the drive currents required for proper operation of the memory element.

As has previously been noted, the hole associated with the interrogate conductor is at right angles to the storage hole. It is this geometrical configuration that makes the BIAX element

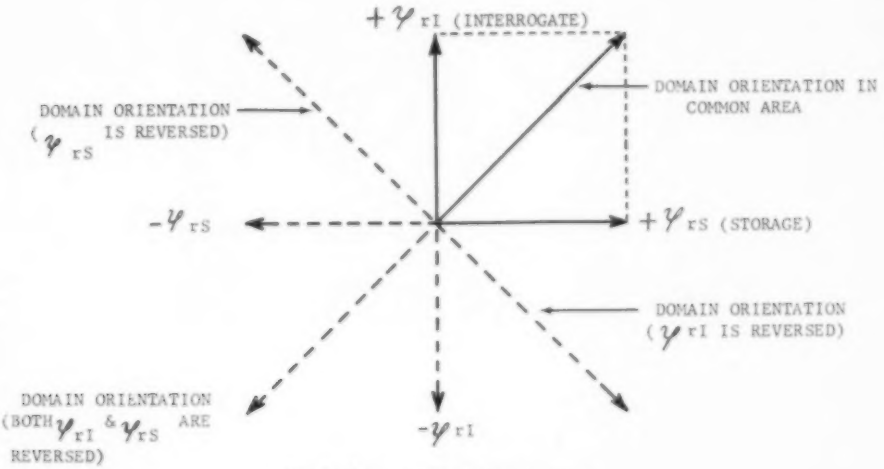


Fig. 8. Domain orientation diagram.

different and superior in performance to other multi-aperture devices having holes whose axes are parallel.

A description of the BIAX read principle is most readily given in terms of the vector diagram shown in Fig. 8. This vector diagram is one in which the domain orientation in the common volume between the two holes is shown in relation to the domain orientation associated with the storage and interrogation axes. Referencing the earlier discussion of this phenomenon, it can be assumed that the magnetic material in this common volume is saturated and the application of currents to the storage and interrogation axes results in a reorientation of the domain vector associated with this common volume. Referring to the particular vector orientation shown in Fig. 8, it will be noted that an increase in interrogate current such as to further increase the flux associated with the interrogate axis will necessarily result in a decrease in the flux associated with the storage axis. On the other hand, an application of the opposite polarity current in the interrogate axis will cause an increase in the flux associated with the storage axis. In either case, the change in the flux around the storage axis will result in an output voltage on the sense winding. The polarity of the sense output voltage for a given system of interrogate will be completely dependent upon the flux orientation around the storage axis with no dependence upon the interrogate current. It should be pointed out that in the interrogate operation, the actual technique involved is a domain rotation phenomena, which, of course, is exceptionally fast and consumes a minimum amount of power. The read operation is nondestructive.

To date, the interrogate operation has been accomplished in approximately 20 mμsec for a BIAX element of dimensions 85 $\times$ 50 mils. As might be expected, the output voltage associated with a given interrogate current is approximately linearly related to the amplitude of the interrogate current. Thus, a 40 mV amplitude signal is normal for a 200 mA turn interrogate current

and an 8 mV signal is approximately normal for a 40 mA turn interrogate current. This can be predicted from considerations of diagrams such as Fig. 8.

To demonstrate that the heating associated with this BIAX interrogation is within practical limits, Fig. 9 presents a stored "1" and a stored "0" BIAX output as a function of interrogate currents at a 10 Mc/s rate. Thus, a complete readout is occurring in $\frac{1}{10}$ of a μsec. Even at these frequencies, no adverse heating effects have been observed. This has been further investigated by experimentally determining the decrease in amplitude of the output signal as a function of the interrogate frequency. The output signal results were then simulated at a very low interrogate rate so that no appreciable internal heating occurs. The ambient temperature of the medium surrounding the element was then increased until the signal output at the low frequency interrogate matched that of the very high frequency interrogate. This determined to a first approximation the internal operating temperature of the element at the higher frequency.

To date, it has not been possible to determine the ultimate frequency at which the BIAX element may be used. It has been observed that the signal output width is at least as narrow as 20 mμsec which is in the vicinity of that associated with thin films. This is also good evidence that the interrogate process is rotational and that very little domain movement other than rotation occurs. Referring again to Fig. 9, it should be pointed out that trace (a) presents the output for a stored "1" while trace (b) presents the output for a stored "0". Photos (e)–(h) are for reduced interrogate current amplitudes. As can be seen in the photographs, the interrogate current pulse is shaped so as to generate a unipolar output.

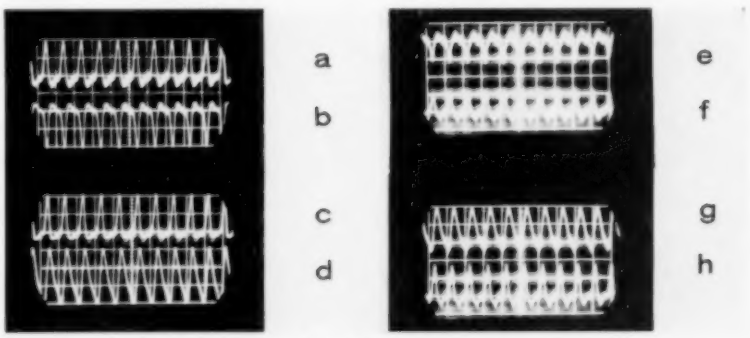


Fig. 9. 10 Mc/s operation.

(a) BIAX "1" output, 40 mV per turn/vert. div.
(b) BIAX "0" output, 40 mV per turn/vert. div.
(c) Flux swing in interrogate hole, 200 mV per turn/vert. div.
(d) Interrogate current, 100 mA-turns/vert. div.
(e) BIAX "0" output, 10 mV per turn/vert. div.
(f) BIAX "1" output, 10 mV per turn/vert. div.
(g) Interrogate current, 20 mA-turns/vert. div.
(h) Flux swing in interrogate hole, 40 mV per turn/vert. div.
Horizontal scale for all traces: 0·1 μsec/div.

The above data on the read characteristics of the BIAX element have employed the concept of a one-pulse interrogate cycle. This system is, of course, only useable for a linear select or word oriented system. However, the BIAX element is also useable in a selection by bit read system. This is essentially the type of system that is used in a coincident current magnetic core memory array. However, the unique method of interrogating the BIAX element allows for a selection by bit read system to be employed where control of the amplitude and width of the selection current is not required.

Referring to Fig. 10, which shows the hysteresis loop associated with the interrogate axis, assume that the operation point of the interrogate axes of all the BIAX elements is that of positive magnetic induction, or B_r. Therefore, the interrogate axes of the BIAX elements will be in the initial state as shown in Fig. 11 (a). The first step of the interrogate operation is to drive the selected X conductor linking through the element to be interrogated into negative saturation or point $-B_r$ as shown in Fig. 10. Referring again to Fig. 11, diagram (b) of this figure demonstrates the new condition associated with the interrogate axis of the storage element. When this first interrogate pulse occurs, there will be a relatively large output signal in the sense conductor which is ignored by the read amplifier. The second pulse associated with interrogate operation will occur on the Y conductor that also links the element to be interrogated. However, the current through this conductor will be positive so as to drive the elements associated with the line into positive magnetic saturation. However, as noticed from Fig. 11 (b), the condition of all the elements associated with this Y-line is positive with the exception of the one to be interrogated. Thus, when this current flows only the element to be interrogated will switch from negative saturation to positive saturation. Therefore, during the second portion of the cycle the output signal obtained on the sense winding associated will be the desired stored information. The third pulse of the interrogate operation will be a pulse on the X-line which drives the remainder of the elements from negative saturation back into positive saturation resulting in the original initial condition. The output that such an operation produces is shown in Fig. 12. In this case, the setup current pulse is increased in duration and decreased in amplitude to exhibit again a very small output during the setup

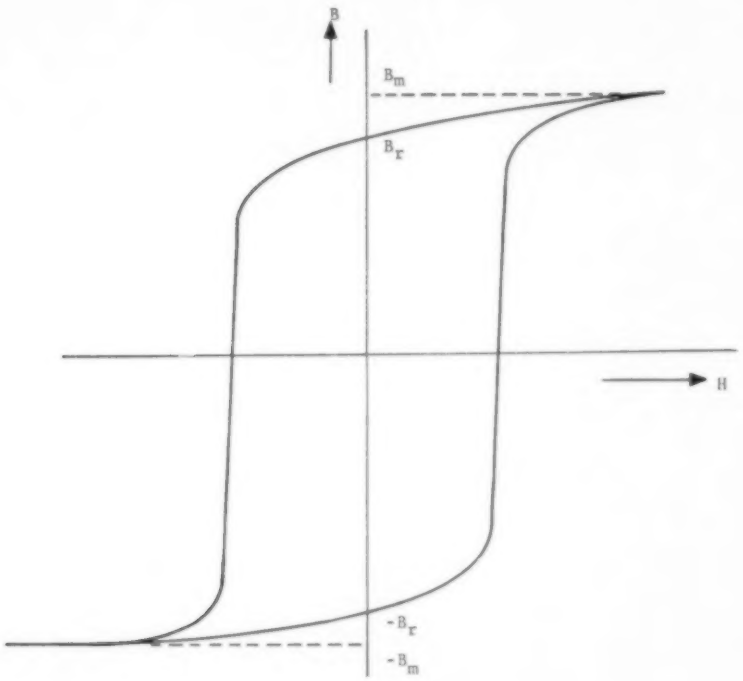


Fig. 10. Interrogate hysteresis loop.

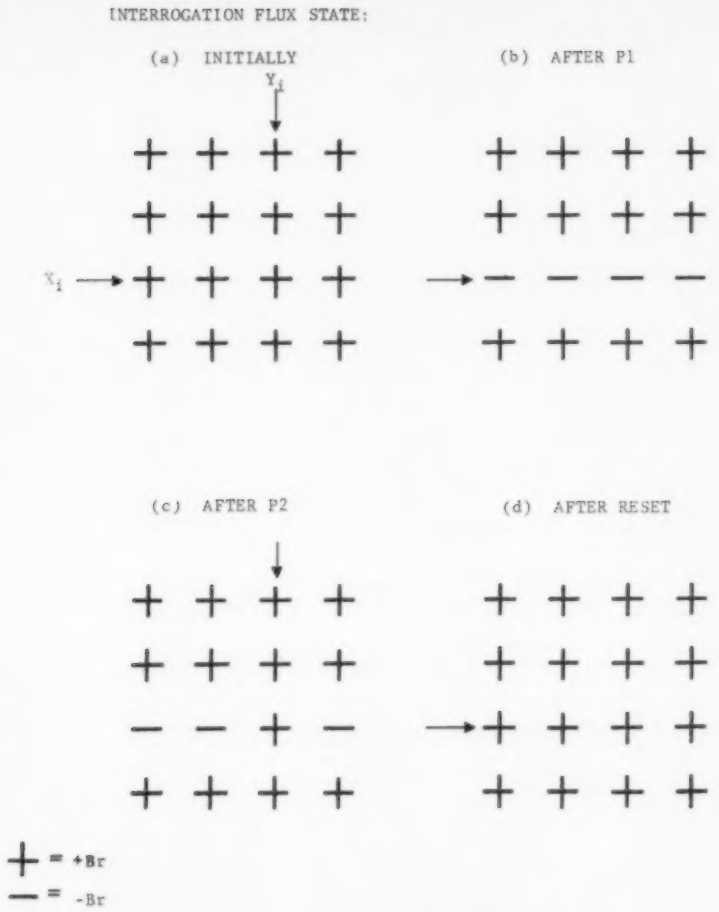


Fig. 11. Two-dimensional (three-pulse) interrogate system.

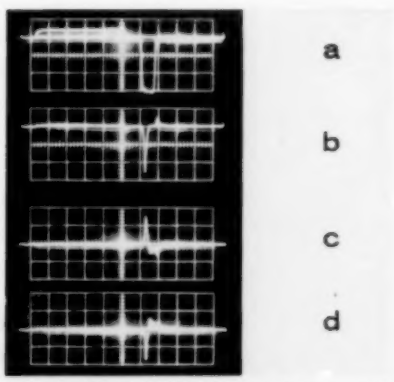


Fig. 12. 500 kc/s two-dimensional interrogation.

(a) Interrogate currents, 0·8 A-turn/vert. div.
(b) Interrogate flux change, 2·5 V per turn/vert. div.
(c) BIAX "0" output, 40 mV per turn/vert. div.
(d) BIAX "0" output, 40 mV per turn/vert. div.
Horizontal scale for all traces: 0·20 μsec/div.

time, whereas the interrogate pulse is of large amplitude and fast rise time, resulting in a large signal output. The signal output for this interrogation system for a stored "1" is shown in the trace (c) and for a stored "0" it is shown in the fourth (d). It should be noted that during this interrogate sequence careful control over pulse amplitude or pulse duration is not necessary since the system is not coincident in time. The disadvantage of this system is that it requires three pulses to accomplish a complete interrogation, whereas the other system requires only one pulse. In addition, the interrogate axis of many elements is switched during the selection process.

A photograph showing the output associated with a stored "1" for this type of interrogate system with identical setup and readout pulses

is shown in Fig. 13. This operation is occurring at a rate of five interrogations per microsecond as can be seen from the photograph (a). The sign of the output associated with the BIAX

Fig. 13. Multi-megacycle interrogation.

(a) BIAX output, 40 mV per turn/vert. div.
(b) Interrogate hole flux swing; 2·5 V per turn/vert. div.
(c) BIAX output, 40 mV per turn/vert. div.
(d) Interrogate hole flux swing, 2·5 V per turn/vert. div.
Horizontal scale for all traces: 0·1 μsec/div.

element is independent of the direction of the flux change in the interrogate axis. The lower photograph demonstrates the same readout procedure at two interrogations per microsecond.

Environment

With conventional core memory techniques difficulty is experienced in attempting to operate over a wide temperature range. This problem is caused by the variation of magnetic material characteristics with temperature and is aggravated when low coercive force materials are used because of their correspondingly lower Curie temperature. As the ambient temperature increases, the ferrite coercive force decreases and the permeability and saturation increase, giving rise to a considerable increase in disturb noise characteristic of the elements. A tendency toward partial switching is also observed which, of course, contributes to unreliable operation. As the temperature increases further (towards the Curie point), the maximum magnetic induction decreases considerably and gradually reduces to zero.

Because of its different principle of operation, the BIAX memory element does not experience these same difficulties. The reason for this is that the read system used for the BIAX memory does not depend upon the existence of a square hysteresis loop for proper operation. The variation of coercivity and the decrease in the sharpness of the knees of the hysteresis loop with temperature rise is not a critical factor in the BIAX memory system. To substantiate this, a series of laboratory tests were conducted in which it was found that the BIAX element performed satisfactorily up to temperatures in the vicinity of the Curie temperature of the material used in the elements. For example, a BIAX array employing square loop materials having a Curie temperature of approximately 180°C was operated for a period of 12 hr at 125°C with no deterioration of its operational characteristics other than a reduction in the output voltage amplitude. This decrease in output resulted, of course, from the fact that the ambient temperature plus some internal heating generated by the three-pulse interrogate system had increased the temperature of the magnetic material to the point where some of the domains lost their magnetic properties by exceeding the Curie temperature. This is what would be expected from normal ferrites since the Curie point is distributed over a temperature range.

The results of these tests are shown in the two photographs which constitute Fig. 14. For the two photographs in Fig. 14, the oscilloscope had a time base of 2 μsec per division and vertical amplitude of 50 mV per turn per division. The time base proceeds from right to left in the photographs. The first large initial pulse at the right is the pulse corresponding to the setting of the interrogation axis. (This information is not used since it is composed of the entire information stored in the X_1 axis.) Near the center of the trace is a positive pulse. This pulse represents the information stored in the (X_1, Y_2) BIAX element and therefore, the output signal. The pulse amplitude is approximately 130 mV. Continuing to the left, a small pulse is present which shows what occurs if the X axis is not driven to the set position before the Y axis is interrogated. This is what would be referred to as a noise pulse if there were such a pulse in this type of a readout system. The temperatures at which these photographs were taken varied

from 25°C, which is shown in trace (a), to 125°C, which is shown in trace (i). It will be noted in the photograph (i) that although the temperature is approaching the Curie temperature of the material, the amplitude of this output is still more than half the amplitude of the signal at 25°C, and that the signal is useable in all respects. Again, in this particular type of operation only the polarity of the signal is critical and not the actual amplitude. Thus, a stored "0" would have the same characteristics except that the polarity would have been negative instead of positive. It should also be pointed out that the data observed in this figure were with respect to a three-pulse interrogate system. With the one-pulse interrogate system, the heating is considerably reduced since only domain rotation takes place. Consequently, the change in output signal is also considerably reduced at elevated temperatures for the single pulse interrogation system.

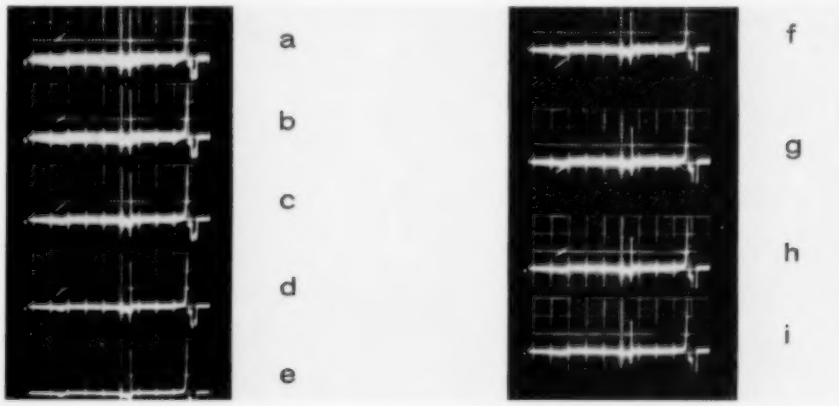


Fig. 14. BIAX temperature characteristics.

(a) Output signal at 25°C (center of trace).
(b) Output signal at 35°C.
(c) Output signal at 50°C.
(d) Output signal at 60°C.
(e) Output signal at 75°C.
(f) Output signal at 90°C.
(g) Output signal at 100°C.
(h) Output signal at 110°C.
(i) Output signal at 125°C.
Horizontal scale: 2 μsec/div.

Material

Because of the read system employed with the BIAX memory element, the material that may be used in the memory element is not critical. Since a high squareness ratio and very sharp knees in the hysteresis loop are not required, a switching type material may be employed in the BIAX memory element as well as a square loop memory material. The switching type material has advantages, of course, since coercivity is less and the area contained in the hysteresis loop is decreased. This, in turn, decreases the power dissipated in the writing process. Various materials from a number of different manufacturers have been employed successfully in the laboratory. In fact, if a read-only type of array is employed, the materials used in the array need not all be of the same material nor even of the same manufacturer. The magnetic material standardized at present for arrays being developed at Aeronutronic is Ferroxcube 6B2 material.

Systems Applications

The inherent characteristics of the BIAX memory element make it extremely well suited

for both military and commercial applications. Its small volume and unique operation make it possible to transistorize the peripheral electronic equipment that is normally associated with a magnetic memory. This is possible because of the low drive current requirements for both write and interrogate operations. In addition, the photograph, the memory package is very small due to the fact that the elements may be packaged in contact with one another. This experimental memory array was used initially as a basis for evaluating these elements in a compact configuration using an access time of 1 μsec.

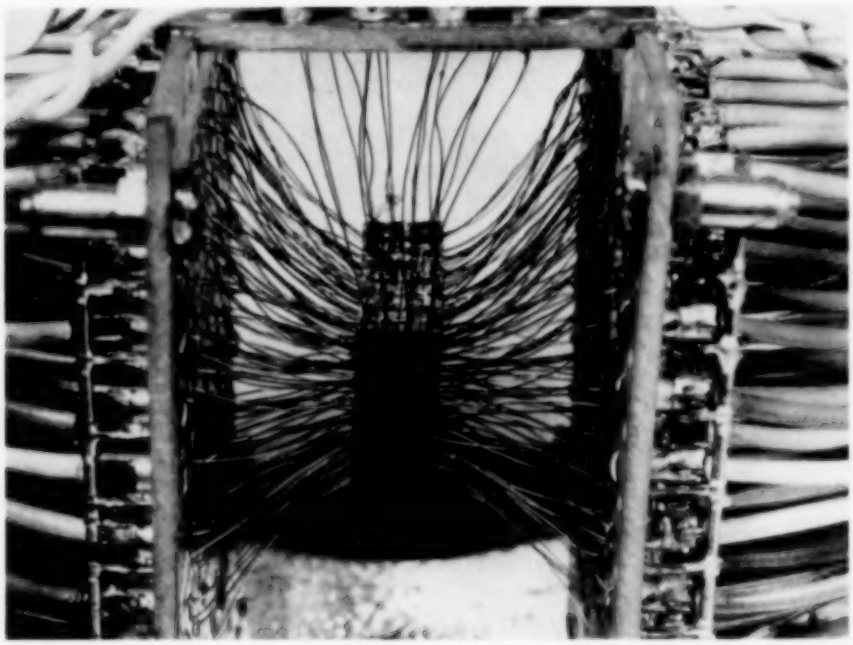


Fig. 15. 256-Bit laboratory array.

the back voltage generated during the interrogate operation is sufficiently low to allow transistorized drive amplifiers to be employed even with relatively long word lengths. This is possible because the normal full write current is approximately 200 mA turns for a 0·4 μsec pulse width, and the corresponding back voltage generated in the interrogate axis is approximately 200 mV per element.

Because of the advantages of the BIAX element, it was placed into production and a series of prototype memories constructed. The first of the memory arrays constructed of production BIAX elements is shown in Fig. 15. This array consists of 256 BIAX elements arranged in an $8 \times 8 \times 4$ array. As noted in

The details of the wiring of the memory array are presented in Fig. 16. As can be seen in this diagram, the elements are wired as a sheet. Wiring other than straight through passes is restricted to a single sheet, thus allowing the wiring of the array to be accomplished in a straightforward manner. It is also noted from Fig. 16 that the interrogate line is perpendicular to the sense conductor. This is, of course, a distinct advantage in reducing the pickup associated with air coupling and has proved very effective in the computer memory system.

BIAX LOGIC ELEMENT

The BIAX element in a slightly altered geometric configuration is used as a logic gating

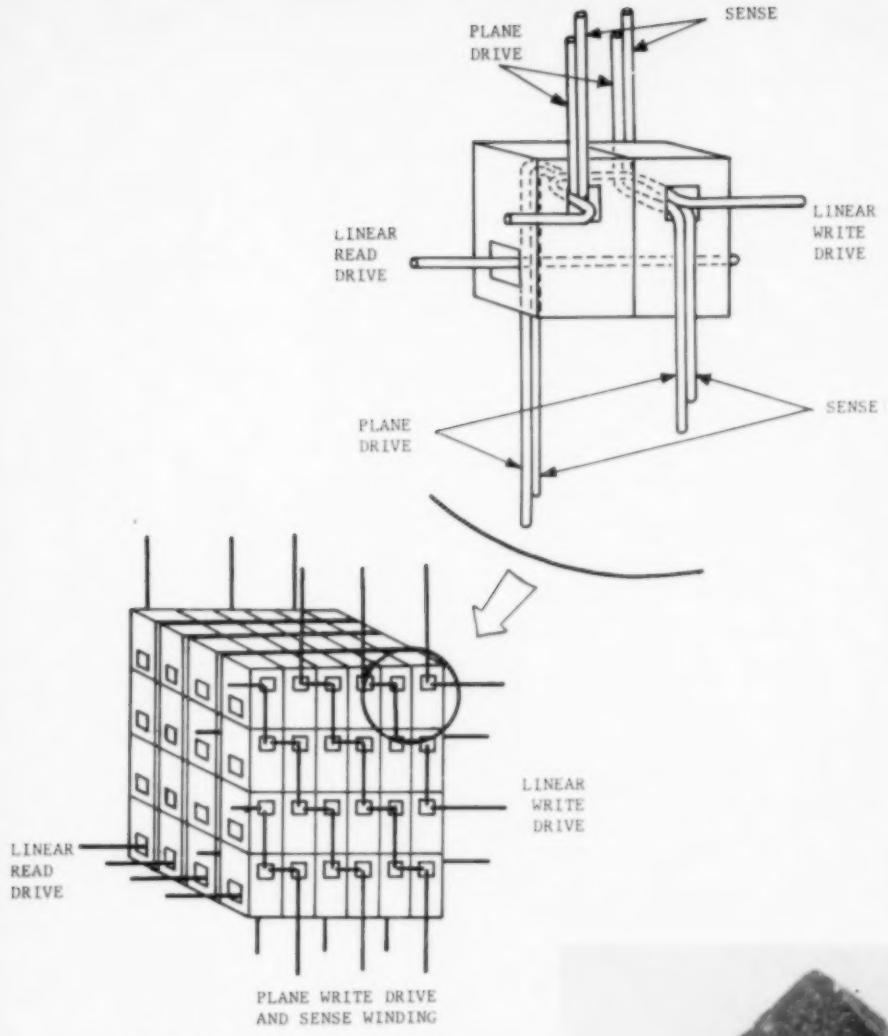


Fig. 16. Memory array wiring.

element and a storage element commonly referred to as a flip-flop. This new configuration makes use of two intersecting orthogonal holes which results in a magnetization–demagnetization type of operation. A photograph of a BIAX logic element is presented in Fig. 17. The operation of the logic element is entirely different than that of the memory element, as will be pointed out in the following discussion.

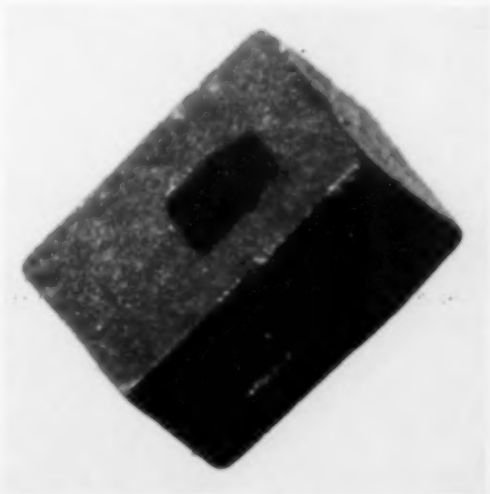


Fig. 17. BIAX logic element.

Basic BIAX Logic Element

In describing the basic logic element, it will

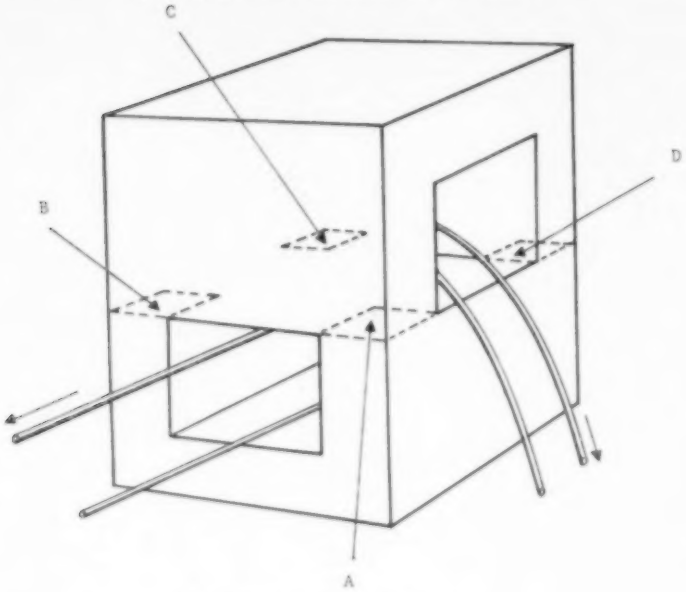


Fig. 18. Basic BIAX logic element.

be assumed that the element is initially demagnetized. Referring to Fig. 18, it will be seen that there are four flux constricting legs; *A*, *B*, *C* and *D*. If a current is introduced through the drive conductor of the lower hole, a flux will be set up in the counter-clockwise direction around this hole. In this process, negligible magnetic flux encloses the conductors associated with the top hole. If after establishing this flux condition, a current pulse is again introduced by means of the lower drive conductor in the direction of the arrow, negligible output would be present on the upper sense conductor. However, if a current of sufficient amplitude is

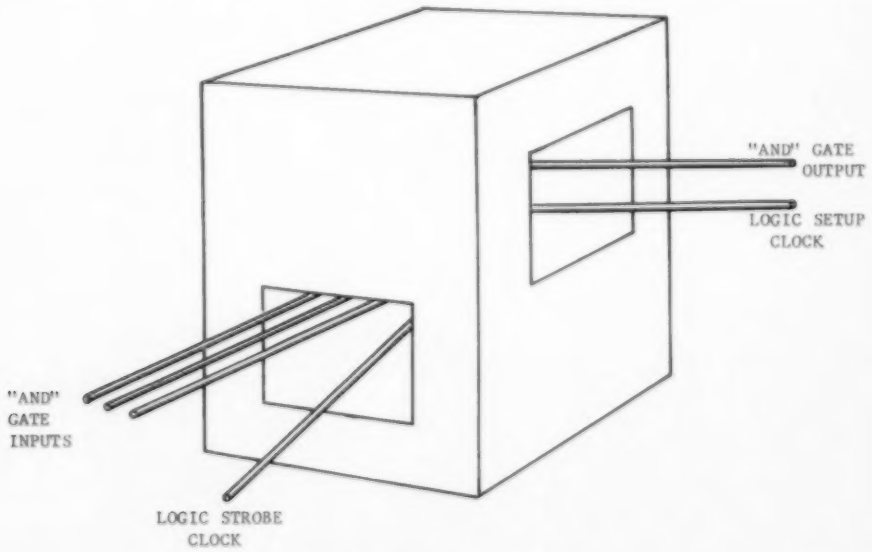
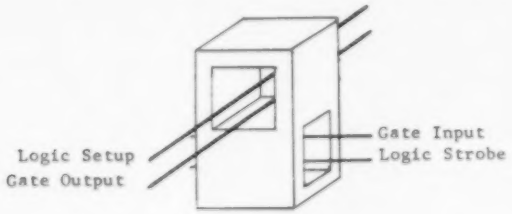


Fig. 19. BIAX logic "and" gate.

introduced through the drive conductor of the top winding in the direction of the arrow, the flux will be set up in the counter-clockwise direction around the upper hole with negligible flux enclosing the lower hole conductors. It

BIAX "and" Gate

The BIAX "and" gate makes use of the element shown in Fig. 18 and the associated wiring is shown in Fig. 19. The input to the gate element from the flip-flop elements is defined as



BIAX Element

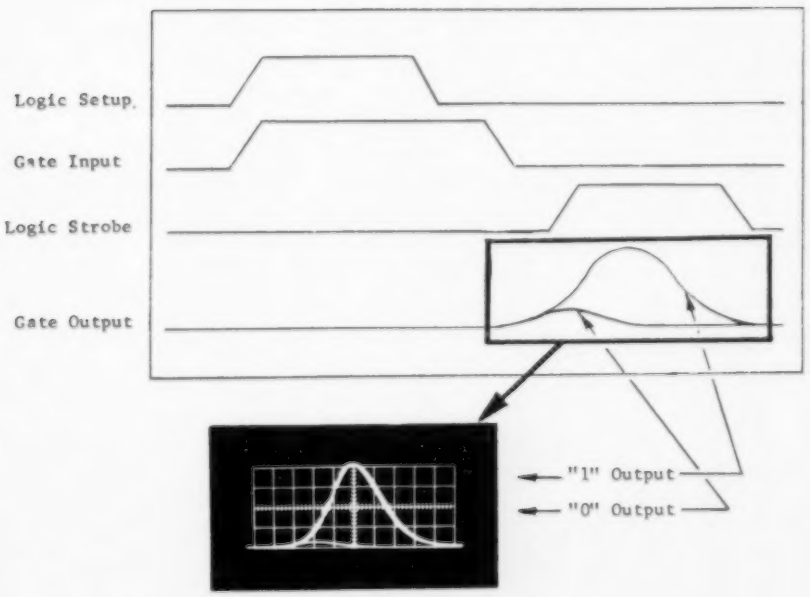


Fig. 20. BIAX gate operation.

might appear that such a switching path is impossible and that switching occurs over the diagonal path (*A–C*). However, experimental laboratory data indicate that only a few per cent, if any, of the total flux switched during this operation is retained around the diagonal path. In addition, fabrication of experimental elements which restrict the diagonal flux paths demonstrated identical operation as the normal BIAX logic elements.

true if no current flows from the associated flip-flop and is false if a current is present. Fig. 20 presents the timing diagram for the logic gating operation.

Consider the case where the element has all gate inputs true. In this condition, when the logic setup pulse occurs, as shown in the first trace of Fig. 20, no gate currents are present. Accordingly, flux switching will occur and the logic setup current will cause the flux to be

established around the upper hole. The signal existing on the output conductor during this setup is ignored. The second pulse that occurs is the logic strobe which takes place in the lower hole. This results in the re-establishment of the flux around the lower conductor, causing a $d\phi/dt$ which gives rise to an output on the gate output conductor, thereby indicating a true gate condition. This is referred to as a "1" output and a photograph of this signal is shown in the lower portion of Fig. 20.

Now, the converse situation will be explained where one or more of the inputs to the gate element from the flip-flop are false. In this case, the initial conditions are as before with the flux established around the lower hole due to the logic strobe current of the previous cycle. As the next logic setup current attempts to initiate flux switching around the upper hole, the gate input currents from one or more of the flip-flops inhibit the switching operation and no flux is established around the upper hole. (The amplitude of the gate current need not be as great as that of the logic setup current since the coercive force of the material assists in inhibiting the switching operation.) Upon introducing the logic strobe, the flux associated with the lower hole is shuttled since switching of this flux to the upper hole did not take place in the previous half of the cycle. Since this shuttle is associated with the lower hole, only a very small amount of noise is established on the gate output conductor. This is a false condition. The output obtained in this condition is called a zero output and is shown expanded in vertical scale in the lower waveform diagram of Fig. 20. A photograph of the efficiency of this process is illustrated in Fig. 21, which shows the signal-to-noise ratios associated with a typical element. It is noted that the signal-to-noise ratio, or the actual true output to false output ratio is approximately 100 : 1.

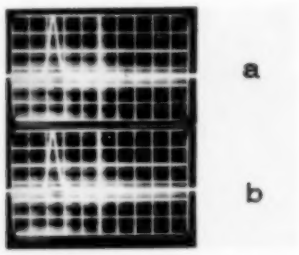


Fig. 21. Signal-to-noise waveforms.

a) Gate output signal and noise on the same sensitivity scale.
(b) Gate output signal with noise expanded by a factor of 4.
Vertical scale: 0·1 V per turn/div.
Horizontal scale: 0·1 μsec/div.

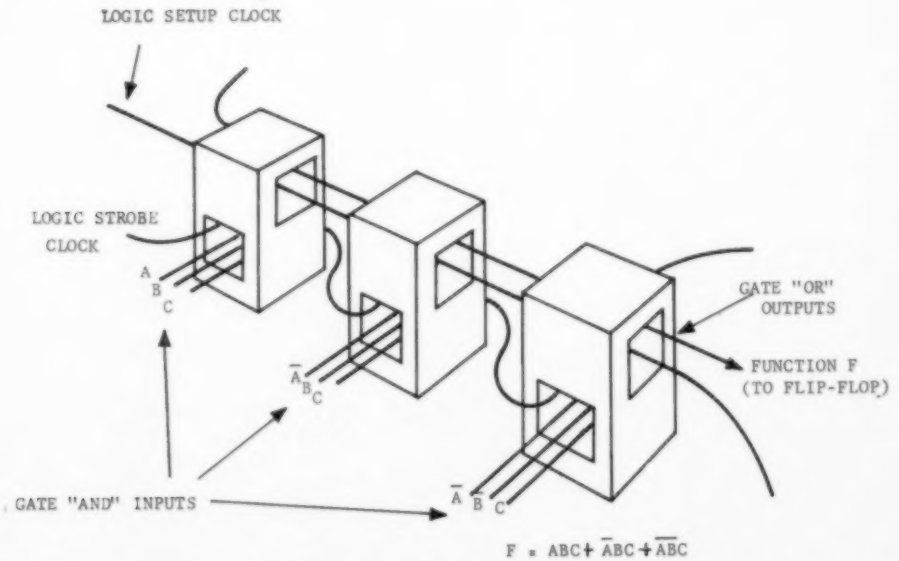


Fig. 22. BIAX logic "or" gate.

BIAX "or" Gate

The actual "or-ing" of "and" gates together to form the "and–or" configuration is a necessity for any computer logical structure. This is easily accomplished in the BIAX gating system as shown in Fig. 22. All that is required is that the "and" output conductor be looped through the various "and" gates which are required to be "or-ed". Because of the exceptionally high signal-to-noise ratio or true-to-false output voltage ratio exhibited by each "and" gate element, "or-ing" large numbers of these elements is possible while still obtaining good system signal-to-noise ratios. For example, assuming each of the "and" gates has a signal-to-noise ratio of 100 : 1, "or-ing" of twenty of these elements together would result in a signal-to-noise ratio of the order of 5 : 1. The "or" conductor terminates as the input to a flip-flop circuit. Flip-flop circuits are composed of a transistor amplifier stage and two BIAX elements which store the function and its complement.

Other BIAX Logic Units

Because the BIAX "and" gate is also a storage element, it will be noted that the "and" output conditions are actually stored in the element during the "and" operation and read out during the strobe clock time. It is, therefore, possible to employ only the BIAX logic element and an active amplifier for the complete logical

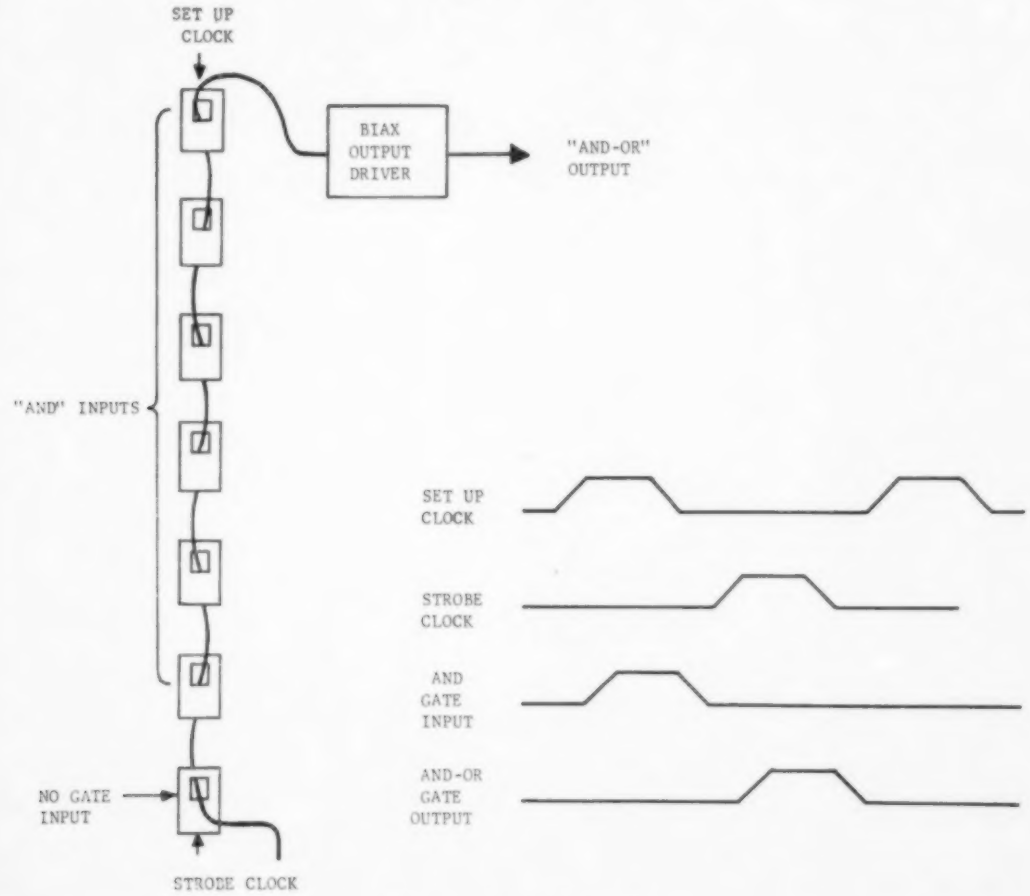


Fig. 23. Basic BIAX logic unit.

structure of a digital system. The amplifier associated with the input to a storage gate is a one-transistor amplifier of simple design. The techniques associated with the use of only logic elements and active amplifiers are well known and will not be considered in any detail here. However, for illustrative purposes, Fig. 23 shows the basic BIAX logic unit and Fig. 24 shows two basic logic units with their associated amplifiers coupled to illustrate a more complex structure. As can be seen from these two figures, the only feedback requirement is that a one-half bit time delay exists between the readout of a given set of logic elements and the input of information into these elements. This may be accomplished as shown in the timing waveforms of Fig. 24.

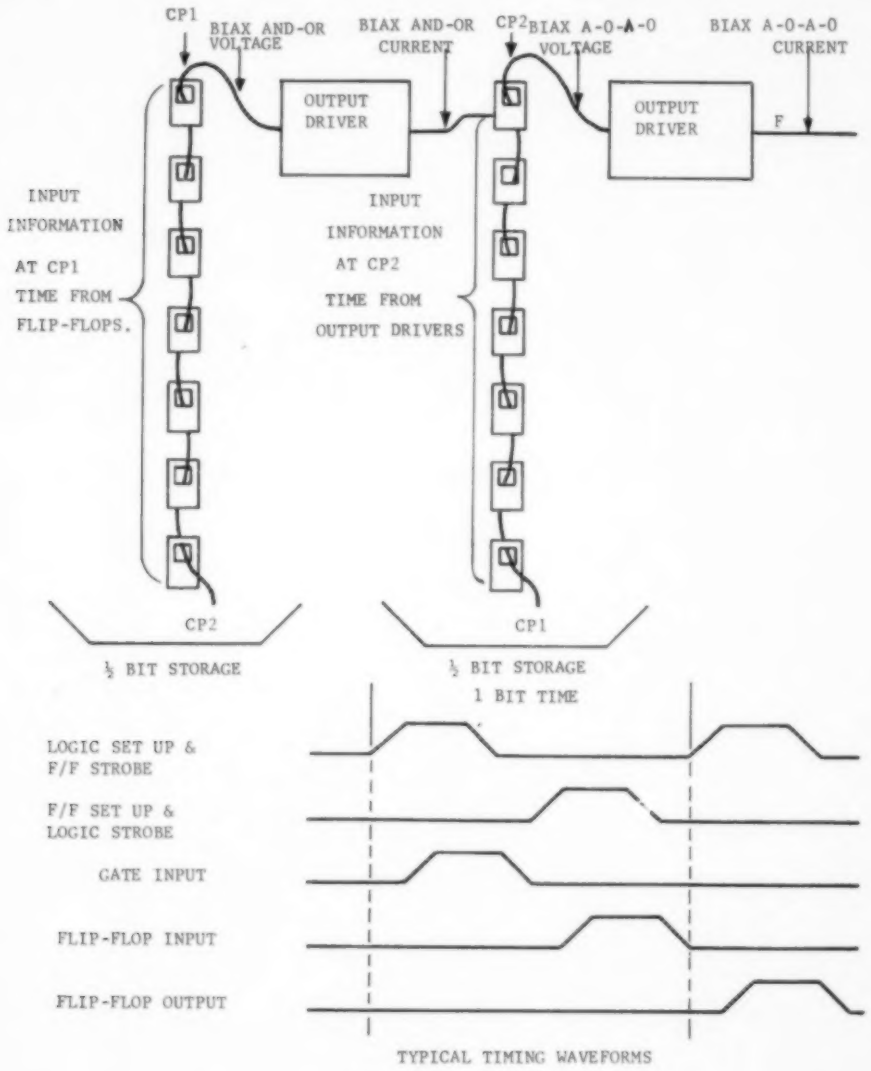


Fig. 24. Cascaded logic.

Advantages of the BIAX Logic Element

As noted in previous sections, the BIAX logic

element has inherent in it many distinct advantages, including:

1. The high signal-to-noise ratio associated with the "and" gate allows "or-ing" of a large number of "and" gates without the problem of poor signal-to-noise ratio in the "or" line.
2. The BIAX logic element is applicable at high speeds due to the small volume of magnetic material employed and the fact that its operation is not critically dependent upon the shape of the hysteresis loop. A $\frac{1}{2}$ μsec cycle time per "and–or" is being used at present.
3. The low drive current requirement of the BIAX logic element makes possible the use of transistors as output drivers. Transistors are also used for the clock sources associated with logic setup and logic strobe pulse generation. (It is of importance to remember that the output drivers only shuttle the BIAX gate elements and, therefore, they do not have to absorb a high back voltage during their operation.)
4. The method employed in the manufacture of these elements has resulted in a low cost element. In addition, each logical "and" gate element replaces approximately six semiconductor gating devices.
5. The logic BIAX element is an exceptionally reliable element since it is a passive ferrite element. Once a given element has been inspected and found to be useable in a system, it is improbable that it will fail except when subjected to an extreme physical environment.
6. The materials employed in the BIAX logic elements are not critical. In particular, a material need not have the high squareness ratio normally employed in logic elements associated with core systems. Almost any ferrite switching type material may be employed in the logic element. The basic material characteristics desired are a low coercive force and as square a material as possible consistent with this low coercivity.
7. Ease of system fabrication is a definite asset of the BIAX logic element as can be seen from a typical wiring system shown in Fig. 25. This diagram shows that the system wiring of the logic elements is extremely simple. In addition, the volume required for this logic structure is very small in comparison with other techniques. As is also shown in the figure, the logic strobe clock is at right angles to the "or" input line and, as such, the normal air coupling is reduced to a minimum during the logic strobe output cycle.

PRACTICAL CONSIDERATION

From the previous discussions, it is evident that the BIAX memory and logic elements have many advantages in a practical digital system. It is these practical considerations coupled with the unique BIAX principle of operation that makes them extremely useful in any digital computer or digital system. The practical advantages of the BIAX element are summarized below.

Packaging

Figure 16 demonstrates the wiring configuration that is used when the BIAX memory element is employed in a word oriented system. As can be seen from this figure, the wiring of these elements takes place in sheet form. During fabrication, the elements are adhered one to another to form sheets and the sheets are wired as indicated in the diagram. This type of construction has proved quite feasible. The spacing between the various planes or plates may be varied as deemed necessary. There is very little heating associated with interrogation and no spacing of the planes is required for satisfactory read-only operation at 10 Mc/s.

The logical structure of a given machine is also very easily packaged as can be seen from Fig. 25. In this diagram, if BIAX elements are placed eight per linear inch, sixty-four "and" gates per square inch can be accommodated. Therefore, a 5×4 in. plane would contain approximately 1,280 BIAX elements which is the equivalent of approximately 7,500 diodes.

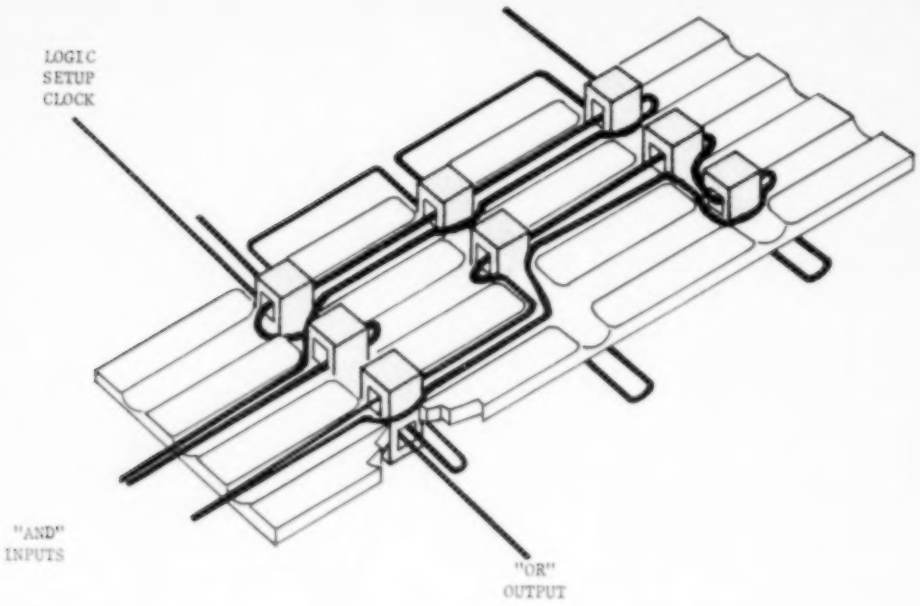


Fig. 25. Wiring of BIAX gate elements.

This is all contained in a $4 \times 5 \times \frac{1}{4}$ in. volume and is sufficient to hold the logic of a fairly sophisticated small general purpose machine. For example, if it is assumed that a particular system incorporating 1,280 "and" gates has approximately 100 set–reset flip-flops associated

Fig. 26. Production BIAX elements.

with it, this would result in a maximum of 200 inputs to the logic board from the flip-flops and 200 outputs from the logic board to the flip-flops. This would result in a maximum of 400 connections between flip-flop storage and the gating structure. The same system would require many more connections if individual semi-conductors were employed.

Production

It should be pointed out that the BIAX memory elements are presently being produced in quantity and that these elements are pressed in one operation and do not require the pressing of one hole with sonic milling of the other. Fig. 26 shows a photograph of a number of production memory elements. Prices associated with the BIAX memory element make it competitive with present magnetic core elements.

The BIAX logic elements are presently being put into prototype production and should be available in production quantities in the very near future. At present, BIAX memory elements are being used with additional sonic milling to make the intersecting hole logic elements. This fabrication technique could be used in production. However, it is probable that the elements will be pressed in one operation as is now being done with the memory elements.

EXTENSIONS OF THE BIAX PRINCIPLE

Double BIAX Memory

The double BIAX memory element makes use of the same phenomena as the normal BIAX memory element. The double element is shown in Fig. 27. The interrogate hole is the center

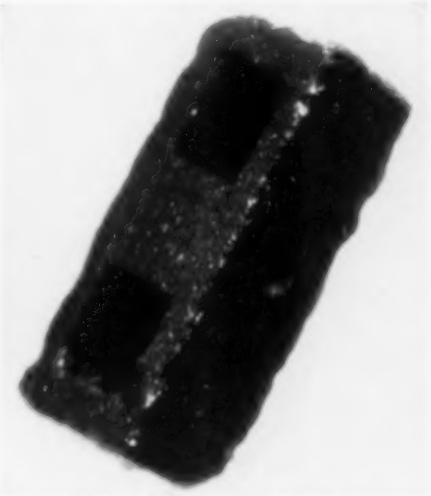


Fig. 27. Double BIAX element.

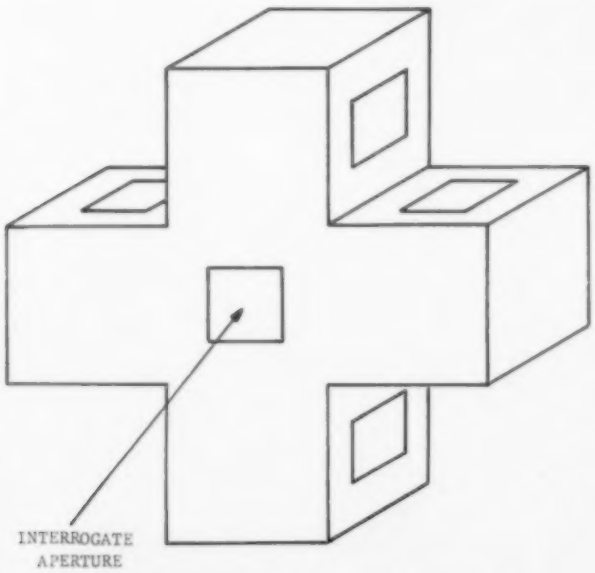


Fig. 28. Four-bit BIAX memory element.

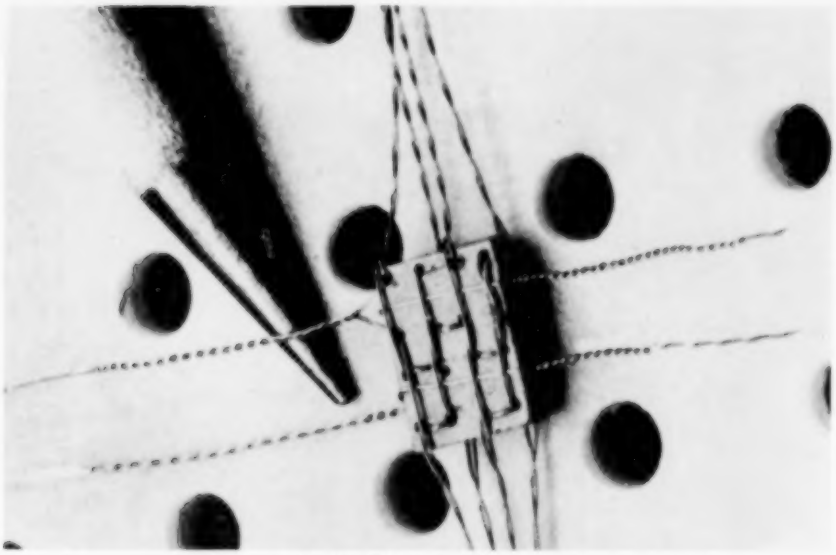


Fig. 29. BIAX memory plate.

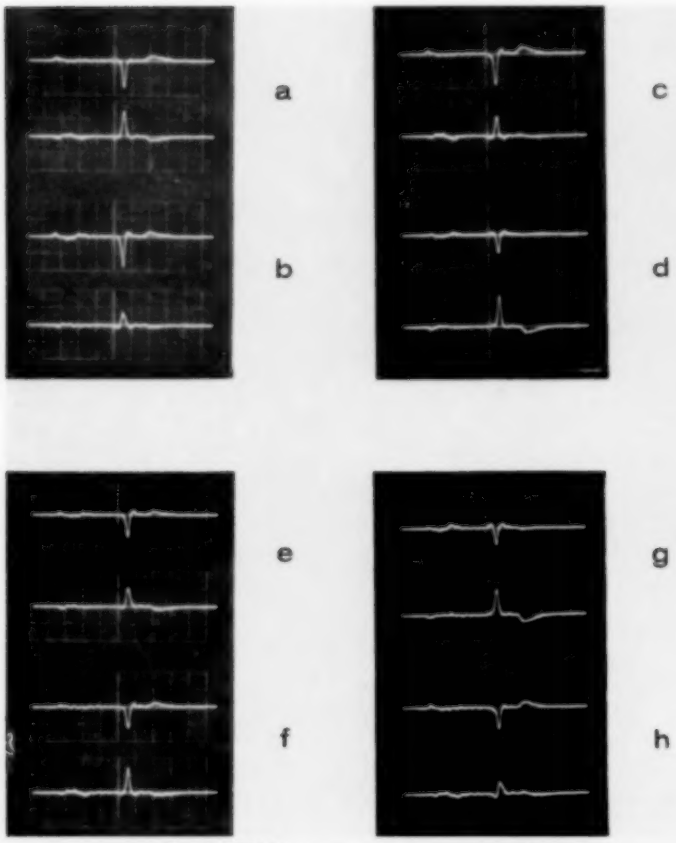


Fig. 30. BIAX memory plate output waveform.

hole; the other two holes are associated with storage of information. It should be emphasized that the use of the double BIAX element permits the interrogation of two binary bits with approximately the same amount of power dissipation as is normally required for a single element.

An even more complex structure is the four-bit unit which is shown in Fig. 28. It should be pointed out, however, that this element would be very difficult to manufacture.

BIAX Memory Plate

A further extension of the BIAX principle is shown in Fig. 29, which is a photograph of a BIAX memory plate. A photograph of the "1" and "0" readout associated with the storage holes of this plate is shown in Fig. 30. This readout employed the three-pulse interrogate system.

TRIAX Element

The BIAX principle can be extended to a TRIAX principle because of the fact that three orthogonal axes exist. Such elements have been fabricated and tested. Typical elements are shown in Fig. 31.

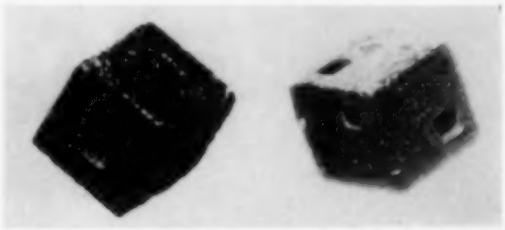


Fig. 31. TRIAX element.

Acknowledgments—The authors wish to acknowledge the important contributions made to this project by W. E. Frady, E. L. Woods and J. Eliades, all of Aeronutronic's Computer Operations.

REFERENCES

1. D. A. Buck and W. J. Frank, *Non-destructive Sensing of Magnetic Cores*, A.I.E.E. Transactions, Part 1, vol. 72 (Jan. 1954).
2. Papoulis, Athanasios, *The Non-destructive Read-out of Magnetic Cores*, I.R.E. Proceedings (August 1954).

THE AIR FORCE APPROACH TO RELIABILITY IN BALLISTIC MISSILE WEAPON SYSTEMS

VINCENT J. BRACHA
Air Force Ballistic Missile Division

Abstract—This paper outlines the Air Force Ballistic Missile Division's concept and approach to reliability. It broadly covers the reliability requirements for a ballistic missile weapon system and the factors involved. It also summarizes the procedures which ballistic missile contractors follow in attaining the required reliability.

INTRODUCTION

The Air Force ballistic missile program encompasses four weapon systems: the Atlas, Titan and Minuteman ICBM's, and the Thor IRBM. These four major projects are managed by the Air Force Ballistic Missile Division (AFBMD) of ARDC, with headquarters in Inglewood, California. Despite the fact that the intercontinental ballistic missile was at the outset (in 1954) largely an engineering unknown, it was recognized at that time that reliability standards of unprecedented stringency would be required.

As applied at AFBMD, reliability is one term in a complex expression that describes what is called "weapon system effectiveness", which also includes factors related to reliability such as logistics, personnel training and efficiency, maintenance procedures, and target condition information. For the purpose of this paper I am only going to discuss reliability. While formalized definitions of numerical reliability do vary with respect to wording, there is an essential agreement that reliability is the probability of satisfactory performance under clearly specified conditions, including time of operation, where appropriate.

The reliability of a ballistic missile system under field conditions is considered from the beginning of countdown through the detonation of the warhead. The system reliability, then, is the reliability of countdown which includes allowances for hold, the reliability of launch, and the reliability of flight which includes warhead separation and detonation within a specified target radius and altitude. The product of these items describes the over-all reliability of the missile system.

It was necessary to define reliability requirements early in the Air Force ballistic missile program because of the extremely short time allowed for completion of the R and D phase. The lead time was less than half of what would normally be allowed for a development project of similar magnitude and complexity. In order to meet these rigorous schedules, the various phases of the design-to-development cycle were pursued concurrently rather than consecutively. This policy of concurrency placed a heavy burden on those responsible for reliability aspects, since it was not possible to proceed in the customary manner through several sequences of designing, testing, evaluating, redesigning, retesting, re-evaluating, and so forth, until the operational design had been thoroughly proven in the laboratory. It should be emphasized that the concurrency policy applied in the Air Force ballistic missile program does not mean a simple compression of the normal sequences. Subsystems go from design directly into production as soon as they cross an established minimum acceptable reliability goal. Production drawings, production tooling and production quality control are used from the outset, even before the first flight test occurs. In this way statistical data, which is the life-blood of any meaningful reliability program, is accumulated over a period of months and years before the system becomes operational.

One of the key elements in the AFBMD

reliability policy is that numerical requirements are being written into contracts at an early stage in the development program. Such a policy is novel in weapon system development programs. In the past, reliability has usually been expressed contractually in qualitative terms, and in few instances has been specified in numerical terms after several years of operational deployment.

ESTABLISHING NUMERICAL RELIABILITY VALUES

In establishing numerical reliability values for the various subsystems and systems which make up the total weapon system, it was recognized that 100% reliability is not only unattainable but generally impractical, even as a design goal. In missile design and development, reliability is in competition with performance, cost, time, logistics and obsolescence. It was also

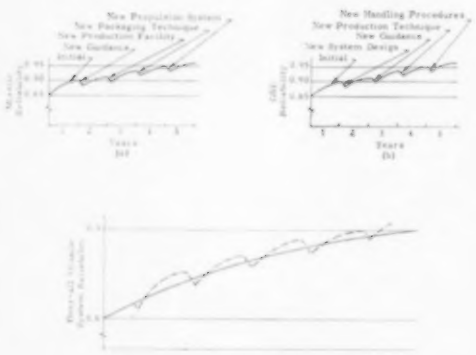


Fig. 1. Typical factors which influence the reliability growth curves of a weapon system.

recognized that there is a growth in reliability. These growth curves usually start at some relatively low value and increase with time, becoming asymptotic to "inherent" values in the ideal case, inherent reliability being the value for which the system is designed and which it is theoretically capable of achieving.

The actual reliability growth curve for a weapon system will not be perfectly smooth but will fluctuate with time because it is, in turn, made up of a number of other growth curves for the various parts of a system. Fig. 1 (a) and (b) illustrates typical growth curves for the missile and the ground support equipment. Identified are some typical major events which can cause fluctuation in the reliability growth but which may ultimately result in higher over-all reliability. Fig. 1 (c) shows the fluctuations in the growth curve of a composite weapon system, such a curve being a combination of the curves of Fig. 1 (a) and (b).

If the Air Force waited until the inherent reliability of the weapon system was attained before accepting it, the system would be, in many instances, close to obsolescence because of the time factor. Consequently, reliability is stated contractually in two terms: a design objective and a minimum acceptable value. The design objective is the inherent value, and the minimum acceptable value is the lowest reliability value acceptable by the Air Force.

ACHIEVING RELIABILITY

One of the most critical systems engineering decisions made in early 1956 was the apportionment of reliability requirements among the various subsystems. Table 1 shows a hypothetical apportionment for an ICBM weapon system. As you can see in this hypothetical example, reliability objectives range from 0·6, as the initial operation capability (IOC) minimum acceptable value, to 0·9 as the ultimate design objectives.

It should be borne in mind that the minimum acceptable figure of 0·6 in the table refers to the over-all weapon system reliability from the beginning of countdown to impact in the target area. Safety factors, even on the earliest operational weapons, are always infinitely higher. The squadron safety officer is able to destroy an operational missile that fails while over friendly territory. A fail-safe arming and fuzing system makes certain that a nuclear warhead will not be accidentally detonated in or over friendly territory, regardless of what happens to the missile. This arming and fuzing system has a reliability on the order of one part in ten million.

In the hypothetical example shown in Table 1, the IOC minimum acceptable values for the

various subsystems range from 0·88 for guidance to 0·99 for the accessory power supply. The apportionment of reliability requirements is made upon consideration of the state of the art, the difficulty of the problem, the contractor's experience in similar projects, cost and time estimates, and other relevant factors. Each contractor then becomes responsible for apportioning the reliability requirements among the various elements of his subsystem.

Table 1. Hypothetical Apportionment for ICBM

Weapon system	Design objective		IOC minimum acceptable value		IOC + 1 year min. acceptable value		IOC + 4 yrs. min. acceptable value	
	0·9		0·6		0·75		0·86	
Airframe	0·994		0·98		0·985		0·99	
Control	0·988		0·95		0·965		0·98	
Propulsion	0·974		0·90		0·930		0·97	
Guidance	0·972		0·88		0·928		0·97	
Ground		0·994		0·97		0·982		0·99
Air		0·985		0·94		0·962		0·98
Computer		0·994		0·97		0·982		0·99
Re-entry Body	0·980		0·91		0·955		0·97	
Vehicle		0·986		0·94		0·965		0·98
Arm, fuzing firing		0·997		0·99		0·995		0·99
Warhead		0·997		0·98		0·995		0·99
ACC power supply	0·999		0·99		0·995		0·99	
Ground system	0·990		0·94		0·970		0·98	

The assignment of reliability goals must be considered to be a useful and necessary effort in that it draws the line between a useful and nonuseful system; however, the problem is in knowing when these goals have been achieved.

Classical statistical theory is inadequate to cope with the small samples and with unique situations which appear in ballistic missile programs. Complete missiles and complex components are so expensive that the word "sample" scarcely applies. An added complication arises from the fact that, in the early development stage, each missile fired is likely to differ from its immediate predecessor in some important engineering characteristic.

At the risk of belaboring some elementary principles of statistical analysis, I would like to elaborate a little on the relation of sample size to reliability confidence. Sample reliability provides a "best estimate" of the *true* population reliability. However, since estimates do vary in precision—the larger the sample size, the greater the precision—it is important that an estimate be accompanied by a statement which describes the precision of the estimate or "how close" the estimate may be to the *true* population reliability. Confidence intervals provide a method of stating the precision of the estimate.

In Fig. 2, which is based on the binominal probability distribution, the solid line represents 90% confidence limits, that is, a 90% probability that the true reliability of the population lies within these limits. To illustrate these concepts, consider a sample size of twenty on the figure. The upper and lower 90% confidence limits are 0·896 and 0·545, respectively. This means that if the sample reliability were 0·75 by actual measurement, i.e. 75% of the sample were successes, the true reliability may actually be something between about 0·5 and 0·9. Note that the confidence curves converge rapidly with the first samples, but that additional testing seems to follow a law of diminishing returns as the sample size is increased. Thus, if the sample size were doubled, i.e. to forty, the 90% confidence limits would still be somewhere between about 0·6 and 0·85. If we double the sample size again, to eighty, our lower confidence limit would only be about 0·65. When these basic statistical principles are applied to flight tests, we see that

an impossible number of tests would be required to verify our 0·75 measurement with a 90% confidence. We would have demonstrated only a 0·7 reliability after 182 flights. Obviously, such an approach would be economically unfeasible and would delay the data at which the weapon could be operationally deployed by many months. Consequently, reliability monitoring, measuring and demonstrating during research and development is based primarily on ground tests using mean-time-between-failures data and other means. Thus sample sizes can be made very large. This entire procedure is in turn backed up by flight data. In addition to this our contractors must continually demonstrate a vigorous and effective reliability program.

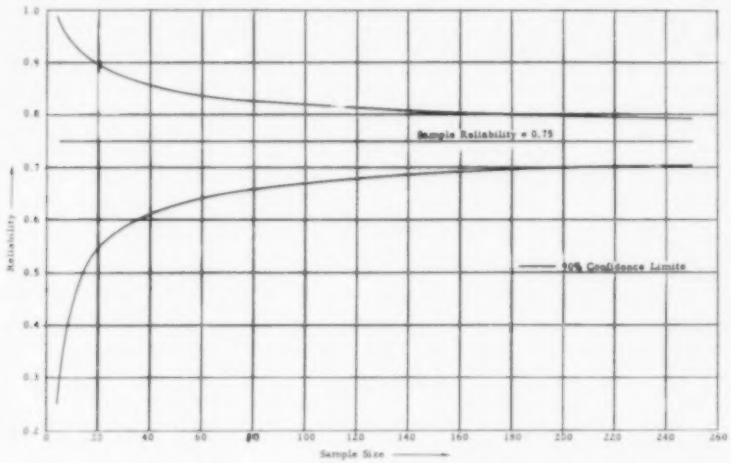


Fig. 2. Confidence limit curves for sample reliability.

A major postulate of the AFBMD test philosophy is that nothing is tested in flight that can be tested on the ground. In this manner large sample sizes are obtained for measuring reliability. The over-all test program can be depicted as a pyramid, as shown in Fig. 3. These tests have dual objectives: first, to show that the parts and the design have the desired performance, singly and in combination, and are compatible with the design: and, second, to show that the life expectancy under operating conditions is such that reliable operation can be expected. Everything from components to captive missile systems is tested and evaluated under carefully controlled laboratory conditions, leaving only performance under free-flight conditions to be proven in flight tests. These ground tests, performed by the contractors at their own or government-furnished facilities, provide a valid basis for reliability measurements. The size of the statistical samples made available by ground tests is truly impressive. Some individual component parts undergo "life" tests of hundreds of millions of hours in a simulated operational

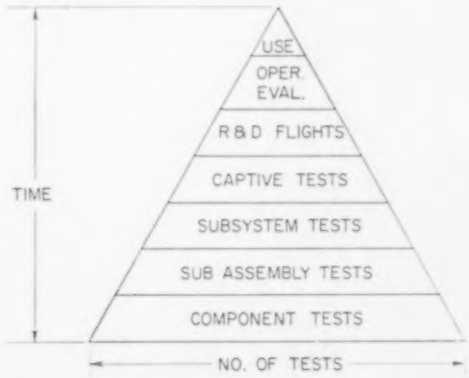


Fig. 3. Test program.

environment. "Black boxes" may be tested for more than 500 hr each. A main engine subsystem may undergo the acceptance tests, which are preceded by something on the order of 100,000 sec of full-duration R and D firings. Finally, captive tests of the complete missile add up to about twice as many testing hours as total flight-test time.

It should be pointed out, however, that test flights do provide useful data on reliability aspects of the weapon system design. Indeed, one of the central postulates of the Air Force ballistic missile test philosophy is that testing for any purpose is also generally testing for reliability. In practice this means that the actual production missiles are used as test vehicles, as contrasted to a policy of developing special purpose vehicles or subsystems to test a function or a principle. Thus the body of statistical data accumulated during the development phase is enlarged, and the costly conversion period that normally intervenes between development and production phases is eliminated.

It must also be remembered that not every subsystem is tested on every flight. The earlier flights are conducted primarily for the purpose of testing the propulsion system and the structural integrity of the airframe. The nose cone may not be flown until well into the flight-test program. Furthermore, on some flights exceptional disturbances or stresses are deliberately introduced in order to give an indication of performance limits. Operational ground support equipment is not used until late in the flight-test program, and the launching of a weapon by military personnel is the culmination of the development phase. The fact that these missiles are heavily instrumented complicates the preparations for launch and multiplies the possibilities of malfunctions. Therefore, test-firing "failures"—as a number of them have been labeled by the press—are in many cases a perfectly normal part of the process of research and development, and do not necessarily indicate anything significant about the eventual operational reliability of the missile. There is no such thing as a complete failure during research and development. Each missile launched, regardless of the outcome, adds to our storehouse of knowledge.

CRITERIA FOR RELIABILITY PROGRAMS

The implementation of the reliability program in Air Force ballistic missile projects begins with the choice of contractors. A special review committee, composed of members of AFBMD and the Ballistic Missile Center (BMC), has developed a list of criteria by which the reliability program of a potential contractor is evaluated. A special *ad hoc* committee establishes the score for each contractor on the basis of the rating which he receives under these criteria. This criteria has been incorporated into AFBMD Exhibit 58-10, "Reliability Program for Ballistic Missile and Space Systems", recently published by AFBMD, which prescribes the basic requirements for establishment of an acceptable reliability program by the Air Force contractor to assure attainment of the reliability goals and/or requirements specified in the contract. Contractors are responsible for determining the applicable portions of the exhibit to be included in subcontracts and purchase orders. The following elements contained in the exhibit are considered to be essential in the development of a reliability program:

1. Concept and Approach

The over-all program should reveal an appreciation of the importance of reliability and of the necessity for an organized approach to this attainment. There should be recognition of the fact that reliability must be inherent in the basic design, that improvement is best accomplished in the early stages of development and test, and that maintenance of reliability extends through the life cycle of a product. The degree of reliability to be attained should be stated in quantitative terms, and the consideration of environmental and operational conditions should be clearly defined. Statements of existing or potential "shortcomings" and explanation of basis for "trade-offs" with other design parameters, such as performance, accuracy, weight, cost, etc., should be made and the progressive steps to be taken to arrive at the reliability

objective should be outlined. These steps or phases should coincide with the monitoring points described below, and should lend themselves to periodic reporting and program analysis.

2. Organization

The Air Force does not presume to tell a contractor how to organize and manage his company. It is recognized that reliability cannot be separated in a practical sense from other aspects of missile engineering. However, the platitude that "reliability is everybody's business", although unquestionably true, should not be accepted as justifying the absence of a clearly defined reliability program and a group of qualified personnel whose specific assignment is reliability. Precautions must be taken to assure that "everybody's business" does not become "nobody's business". Consequently, the reliability organization within the over-all management organization should be such that it reflects an important position in the management effort. The reliability organization must include a single group, either in line or staff, with assigned authority of specific responsibilities to cover engineering, manufacturing, quality control and purchasing functions. The group must be able to integrate these functions into an effective program. There should be a director or manager to provide the necessary co-ordination, not only at the policy setting stage, but, more importantly, in the operating stage so that the program is kept on the policy course and not allowed to drift off to the detriment of any one aspect. The director, or manager, should be responsible and have the authority for ensuring that the necessary services (data processing, drafting, training, etc.) are provided as needed by the program and for seeing that the executive offices are informed of and in accord with all aspects of the reliability program.

3. Programming

There should be a time-phase plan covering each point of the reliability program. The authority and responsibility of the head of each working group should be clearly specified so that specific reliability activities can be directed and their accomplishment measured, both as to progress and completion. All reliability activities should be scheduled so that they are properly phased where phasing is required, and so that they may be accomplished when required. The program should be so well defined, and broken down to such detail, that the people, skills, facilities and elapsed-time requirements for each specified activity can be estimated with sufficient effectiveness to establish program cost. The program must include a number of fixed points at which reliability can be objectively reviewed. The sequence of program monitoring points given below has been taken from the Department of Defense's "Proposed Reliability Monitoring Program for Use in the Design, Development and Production of Guided Missile Weapon Systems". These points are not intended to delineate an ideal guided missile life cycle, but to suggest qualitative points at which the program can be monitored. Some of these points will not exist in many guided missile life cycles, and in some cases the points will be telescoped or overlapped. Specific reliability monitoring points will be negotiated with each contractor. These points are as follows:

(a) Detailed design study
(b) Preprototype
(c) Prototype
(d) Preproduction demonstration
(e) Demonstration of service readiness
(f) Service evaluation
(g) Full-scale production
(h) Demonstration of major product improvement

4. Failure Reporting System

A critical element in the AFBMD reliability program is the failure-reporting, analysis and feedback system. Each contractor is required to have a system for recording all failures and malfunctions that occur in-plant and all those which occur at the test site. The system must include records of accumulated operation time or operation cycles on specified items. These data are analysed and transmitted to the design

or manufacturing authority, whichever is appropriate, and to the other specified agencies. The reported in-plant data is classified according to the relative significance of these failures with respect to the achieved reliability. Provisions must be included which will assure that effective corrective actions are taken to prevent repetition of the failures.

5. Vendor Selection and Control Program

The contractor is required to maintain a supplier selection program, based upon review of the supplier's quality control system and examination of his facilities, past performance, etc., to assure that the suppliers are capable of maintaining the required level of performance. Records of each supplier's performance is maintained and an approved source list established and reviewed periodically. The contractor must maintain adequate channels of communication with all suppliers, and furnish each supplier with adequate specifications, acceptance criteria, information and contractual requirements for reliability of the supplier's products. The contractor must exercise proper surveillance over the supplier to assure satisfactory performance, assist him in problem solution and provide him with feedback for corrective action as necessary.

6. Environmental Testing

The reliability program must include functional and environmental testing of design in order to provide for feedback of information to assist in the improvement of the inherent reliability of design and serve as a means of measuring reliability. There should be a realization of the necessity to qualify fully all systems, subsystems and equipment before production in quantity is undertaken. This applies to CFE as well as GFAE, and is not limited to those items whose specifications call for qualification testing for formal inclusion on the qualified products list. This concept requires a time-phased scheduling of qualification testing to ensure completion by the time the first production article comes into being.

7. Quality Control

Contractors are required to have a quality control program which will provide assurance that the reliability of each item is not degraded in the manufacturing process. This program must take into consideration the proficiency of the production workers; training programs must be undertaken where the level of proficiency is too low. In addition, specifications and standards must be established for use in shop acceptance, source and receiving acceptance, and reliability measurement. All defects must be classified according to their importance in a reliability sense. The classification of defects is prepared so as to permit the acceptance inspection at the most advanced manufacturing level possible, with as few separate intermediate inspection steps as possible. There should be recognition that, while the function of quality control is concerned primarily with determining compliance with existing specifications and acceptance criteria, this activity can also contribute substantially to reliability by feedback to and close liaison with engineering, manufacturing and purchasing elements of the organization. This requires that the organizational status of quality control have equal access to the ear of management at the reliability policy level and partnership participation at the working level. In operational procedure, there should be evidence that quality control is ensuring compliance with Air Force Standards (MIL-Q-5923), also that it is provided with and utilizing those higher standards and criteria which experience and "state-of-the-art" have indicated will contribute to improved reliability.

8. Reliability Studies

There should be full appreciation that reliability demands quantitative treatment and that such treatment involves advance study—that it is not sufficient to wait for failure-reporting to set the pattern. Such studies should be aimed at establishing quantitative requirements to include specification of required function (performance limits), operating time and environment as well as the required reliability. The requirements should be realistic,

determined as early as possible and revised as necessary. Environmental conditions specified should cover all phases from factory to target including shipping, storage, handling and flight. Since the reliability of a weapon system is a function of all the elements of the system, the reliability requirements for subsystems, equipments and parts must conform to the over-all system requirements. The terms used to express requirements must be meaningful from the standpoint of mission success and flight safety.

9. Personnel Indoctrination

The contractor's reliability program must contain provisions to indoctrinate all employees whose work contributes to the reliability of the product so that they are aware of their contributions. Contractors should enhance education and training in all aspects of reliability for all personnel, including management. There should be a continuous dissemination of reliability information through poster campaigns, plant newspapers and other media in order to inspire reliability consciousness in all employees. Reliability personnel should be encouraged to prepare and deliver papers on reliability at national symposiums and technical meetings.

10. Reliability Demonstration

The contractor must demonstrate quantitatively the minimum acceptable reliability in a manner that is mutually acceptable to both the contractor and the Air Force. As pointed out in Fig. 2, statistical techniques used in demonstrating reliability depend upon fairly large test samples from production lots of items which are essentially alike. It is often difficult, if not impossible, to obtain such samples of ballistic missiles or major components. AFBMD does not expect contractors to perform miracles. However, contractors are required to demonstrate, to their maximum capability, the attainment of reliability objectives. Methods of demonstration are negotiated between the contractor and AFBMD. Oftentimes the method of estimating whether or not a component or subsystem has achieved satisfactory reliability consists of assuming an exponential distribution of time-to-failure and of determining the mean-time-to-failure of the item. This is then converted to a reliability figure. At other times, where the cost of testing a sufficient number of samples is prohibitive, the demonstration of reliability may take the form of a strong and effective reliability program.

MONITORING RELIABILITY PROGRAMS

Having specified the basic elements of an acceptable reliability program contractually, AFBMD, in conjunction with the Ballistic Missile Center (BMC) and Space Technology Laboratories, Inc. (STL), can more easily monitor the contractor's reliability efforts during the performance of the contract. Specific monitoring and technical guidance is done by the reliability department of STL with the help of Air Force quality-control plant personnel. STL was hired by the Air Force to assist AFBMD in the systems engineering and technical direction of all ballistic missile projects.

SUMMARY

Now, in summary, I hope that in speaking in such positive terms about AFBMD's reliability philosophy and requirements I have not left the impression that we feel that we know all the answers. Reliability is a discipline that is still in its infancy. We feel that our approach is based on sound engineering and statistical principles, but our function is to provide guidance rather than dictation to our contractors. We are striving to develop procedures which will be mutually acceptable to AFBMD and the contractors, and which will result in the most effective weapon systems.

We recognize that, in missile design and development, reliability is in competition with performance, cost, time, logistics and obsolescence. We have not found a general objective analytical formula for arriving at an optimum balance among these trade-off functions; this must be a matter for executive judgment. We

feel that it is unrealistic to expect trade-offs to always be in favor of reliability. Therefore, we do not expect complete failure-free operation from complex weapons. 100% reliability is not only unattainable but generally impractical, even as a design goal. However, we do expect our contractors to carry out an extensive and effective reliability program and to demonstrate quantitatively the attainment of reliability objectives to the maximum extent.

RELIABILITY EVALUATION OF LARGE SOLID ROCKET ENGINES DURING ENGINEERING DEVELOPMENT

M. LIPOW and D. K. LLOYD*
Space Technology Laboratories, Inc.

Abstract—A method of evaluating reliability of large-scale rocket engines during a research and development program prior to missile flight tests is presented. Reliability estimates are obtained even though the configuration of the engine is undergoing change and the objectives of test firings vary. Each engine is apportioned into principal subsystems which are screened for their degree of representation of the final configuration. The intention of each test firing is determined prior to the test and the behavior of the principal subsystem when tested within the engine environment is classified according to pre-specified ground rules as a success, failure or exclusion. These results are then statistically combined to give an estimate of engine reliability.

FOREWORD

The concepts used and the approach taken to solve the particular problem discussed in this paper may be generalized to apply and be utilized for systems in which the following situations occur:

(1) It is desired to evaluate the reliability of the system against its end use.

(2) The system is in a state of continuous development towards an end configuration.

(3) The intentions of the tests vary as the system evolves and there are no or few tests specifically for reliability evaluation.

(4) There is a limited number of systems available for testing and/or there is a limited number of times (perhaps only once) that a system can be tested.

For large solid rocket engines, each of these situations occur. The specific details of the proposed solutions are presented after the practical and statistical problems are discussed. It should be noted that this paper specifies how the reliability of the system can be evaluated; it does not discuss how reliability may be improved.

OBJECTIVES AND PROBLEMS

The objectives are to define and measure during a research and development program characteristics which are indicative of the engine's ability to perform reliably its function of propelling a missile of one or more stages into a certain target area. To accomplish these objectives three major practical and statistical problems must first be considered:

(a) Relating engine reliability requirements in an R and D program to the weapon system's operational reliability requirements.

(b) Obtaining valid reliability data in an R and D program when the objectives of the tests vary and the configuration of the engine undergoes continuous modification.

(c) Making efficient reliability estimates based upon the results of a limited number of full-scale engine test firings.

DISCUSSION OF PROBLEMS

(a) The ultimate intention for a given missile is that it will fall within a certain target area. This intention can be projected into single-stage engine reliability requirements as implying that for successful operation of the missile there should be no catastrophic failure which would result in mission abort nor any unsatisfactory performance which would result in a significant target miss. On the other hand, operation of a given engine which exhibits no catastrophic failure and for which the

* Work performed under Air Force Contract AF 04(647)-309.

performance parameters all lie within specified limits would, assuming successful operation of the remainder of the missile, produce a scatter of hits in and around the target area. The ground rules which interpret the success or failure criteria of the weapon system into the success or failure criteria of the engine may, to a certain extent, be based on engineering and arbitrary judgments. However, providing the ground rules are clearly stated, understood and rigorously applied, reliability estimates can be made which are valid within the framework of those ground rules. Furthermore, only developmental test stand firings can be utilized for reliability evaluation during the program, since flight tests will not yet have been conducted. Thus, the effect of performance interactions of an engine with the remainder of the systems in the missile cannot be comprehensively known during this period. Therefore, if reliability is defined as the probability of a successful operation of the weapon system, then the reliability of the engine, as a functioning subsystem of that weapon system can be interpreted as the probability of successful operation of the engine. This in turn can be regarded as the probability of performing within engine specifications. Thus, the relationship between the weapon system and engine is defined by the specifications.

The numerical value of reliability which the engine may be expected to display or contractually demonstrate would depend on the "state of the art" of the engine, the number of units available for testing, and the reliability requirement of the weapon system.

(b) The applicability of engineering development tests for reliability evaluation is dependent on the validity of the data and the intention and circumstances of the tests generating the data. Generally, the nature of developmental testing differs from reliability testing. Reliability testing would usually involve a large homogeneous sample of engines representative of the final configuration. It is generally not practical to produce a large number of expensive engines specifically to demonstrate the system's reliability. Even if this were done, by the time production and testing were completed the reliability results would refer to an obsolescent configuration, since simultaneous engineering testing would probably have resulted in further design improvements. Consequently, in this program reliability must be estimated from the results of R and D engineering tests. However, engineering testing is usually performed with small sample sizes on changing configurations. Therefore, it is necessary to take into account the objectives of an engineering test, to ascertain how these objectives differ from those of a reliability test and to see if it is possible to reconcile the two. This must be done without restricting the exploratory nature of development testing.

This is accomplished by "screening" the engineering tests for reliability use by means of the "Declaration Policy" and determining the degree of representation of the engine towards its flight configuration. Thus, even though the item is evolving during the R and D test period, representative data can be obtained and used to determine a valid estimate of reliability.

(c) The foregoing discussion implies that only full-scale engine tests will be used for reliability evaluation. A development program for state-of-the-art solid propellant engines involves a great deal of experimental testing with subscale engines. Initial feasibility studies for new or modified propellant formulations are best and most economically undertaken in small and subscale engines in order to provide evidence that the propellant will satisfy internal ballistic requirements in the full-scale engines. In many instances, accurate scaling predictions of internal

ballistic performance can be established. Further testing with subscale engines is undertaken in order to determine properties of charge and case designs, insulation materials, movable nozzle designs, etc., for the full-scale engine.

However, there exist differences between subscale and full-scale engines which cannot be completely resolved by the use of scaling or correlation factors. For example, these differences might be due to the lack of sufficient knowledge of the mechanical properties of scaled-up propellant charges, such as of stress magnitudes and patterns, which are intimately related to problems of propellant cracking and propellant-liner separation.

Thus, while the engineering information obtained from subscale engine tests is essential for the development of performance and reliability of the full-scale engine, the best evidence to ascertain that the end product requirements will be met is obtainable only from tests of full-scale engines sufficiently representative of final design configuration.

Since there are only a limited number of full-scale engines representative of final configuration available for testing, the statistical method of making reliability evaluations becomes of great importance. The statistical technique, introduced in this program, supersedes "the product rule" method by allowing more efficient confidence limits to be estimated.

DESCRIPTION AND CONDUCT OF THE PROGRAM

The program is divided into four parts:

(a) A method of system apportionment;

(b) a method of classifying R and D test results for reliability evaluation;

(c) a declaration policy;

(d) a new technique of estimating reliability.

System Apportionment

To compensate for the relatively small number of tests of full-scale engines, it is essential that all representative data be utilized. This is done by apportioning the engine into three principal subsystems, and each engine test is evaluated in terms of the behavior of the principal subsystems. Thus, the fact that an engine is not fully representative of the final configuration in any one test will not prevent the evaluation of those subsystems which are operating in a configuration or manner representative of flight status. The three principal subsystems are:

The propellant charge-ignition subsystem (A), the case-liner-internal insulation subsystem (B), the thrust vector control subsystem (C).

Engine reliability estimates are made from the principal subsystem test results (tested within the environment of the full-scale engine) and can begin with the first test firing.

Applicability of Principal Subsystem

The principal subsystems tested in a full-scale engine firing will be classified as applicable or inapplicable for purposes of reliability evaluation. In order to determine which of the subsystems being tested in any full-scale test firing are sufficiently representative of flight configuration to be useful for reliability evaluation, it is necessary to set up criteria which determine the subsystem's applicability (cf. Table 1).

Test Result Classification

Those tests which are applicable will be evaluated and categorized as exclusion, failure or success. A result may be excluded *prior* to this test, the circumstances which permit this are listed in Table 2.

Each principal subsystem will be classified as having succeeded or failed depending on whether its performance in operational use would have resulted in a successful or failing flight. The exception to this would occur when for causes external to the subsystem, the subsystem was not given the opportunity to succeed or fail. The detailed ground rules are given in Table 3.

Table 1.

Principal subsystem	Required characteristics for applicability
Subsystem *A*: Propellant charge-ignition	W_p = propellant charge weight, lies within $\pm 1\%$ of model specification limits.
Subsystem *B*: Case-liner-internal insulation	(1) Engine configuration includes flight-weight case and end closure as specified by current weight and balance status report. (2) Propellant weight as specified for subsystem *A*.
Subsystem *C*: Thrust vector control	(1) Engine configuration includes flight-weight movable nozzles. (2) Flight-weight APS actuator subassembly. (3) Propellant weight as specified for subsystem *A*. (4) Nozzles must be intended to actuate during test firing. (5) Predicted action time not less than model specification limit unless thrust termination operation is being tested.

Table 2.

	All principal subsystems (*A*, *B*, *C*)
Pretest exclusion (subject to approval of prime contractor/program manager)	(1) The principal subsystem does not have required characteristics for applicability. (2) Due to the intention or circumstances of the test, a particular subsystem may be excluded due to stated uncertain characteristics of operation relating to performance and/or possibility of malfunction; e.g., if inspection shows propellant voids or cracks of a sufficient degree so that malfunction would be expected; then the propellant charge-ignition subsystem may be considered for exclusion. (3) If internal ballistic performance is predicted to be outside current model specification limits, but the test firing is approved, then performance will be excluded. A maximum of two exclusions under this ground rule prior to the formal PFRT test program is allowable. No exclusions under this ground rule are allowable during the formal PFRT program. (4) Provisional exclusion: A subsystem will be excluded if there is a failure of an experimental part which has been so listed on the declaration form prior to the test, and which is being tested for the first time in a full-scale engine. However, if the subsystem fails due to the failure of a nonlisted part, it will be classified as a failure. (5) Provisional exclusion: A subsystem will be excluded if there is a failure of an obsolete part which has been so listed on the declaration form prior to the test. An obsolete part is defined as a part used in a test configuration for reasons of expediency, but for which there already exists a reliability design change* for that part.

Note: Reasons for all exclusions must appear on the declaration form prior to the tests.

* A reliability design change is defined as a modification to correct a previously observed failure, and must be intended to appear on all subsequent engines.

Table 3. Ground Rules for Classification of Test Results

	Propellant charge-ignition subsystem (*A*)
Success	(1) The subsystem does not fail and is not excluded; and performance is within the current model specification limits. (2) It operates without failure with performance outside current Model specification limits, provided that this intention is so stated on the declaration form, and approved, prior to test. (See limitation in Table 2 (3).)
Failure	(1) The igniter fails to operate or fails to ignite the propellant. (2) The ignition delay is greater than the maximum values specified by the current model specification limits. (3) The ignition peak pressure/thrust is greater than the equilibrium chamber pressure/thrust. (4) Rough combustion; i.e., chamber pressure/thrust peaks $>10\%$ of equilibrium chamber pressure/thrust. (5) Engine blow-up attributable to propellant or igniter. (6) The performance values lie outside the current model specification limits when the performances were declared to be within the current model specification limits.
Post-Test Exclusion	(1) Tests which fail due to causes external to subsystem *A*; e.g., other principal subsystem failures, test operator error, instrumentation or facility malfunction, provided that failure of subsystem *A* had not already occurred. (2) Failure of the subsystem due to the failure of an obsolete or experimental part so listed prior to the test on the declaration form. (3) Tests in which subsystem *A* did not fail but did not have the opportunity to satisfy the declared intention of the test.
	Case-liner-internal insulation subsystem (*B*)
Success	(1) The subsystem does not fail and is not excluded.
Failure	(1) Case or end-closure burn-through during normal operation.
Post-Test Exclusion	(1) Tests which fail due to causes external to subsystem *B*; e.g., other principal subsystem failures, test operator error, instrumentation or facility malfunction, provided that failure of subsystem *B* had not already occurred. (2) Failure of the subsystem due to the failure of an obsolete or experimental part so listed prior to the test on the declaration form. (3) Tests in which subsystem *B* did not fail but did not have the opportunity to satisfy the declared intention of the test.
	Thrust vector control subsystem (*C*)
Success	(1) The subsystem does not fail, is not excluded and performs as programmed for the duration of the test.

Table 3—*Contd.*

	Thrust vector control subsystem (*C*)—*Contd.*
Failure	(1) Movable nozzles stick, jam or do not deflect as programmed. (2) Gas leakage or burn-through in any part of the movable nozzle assembly occurs. (3) Nozzles under- or over-travel intended deflection > 0·5 degrees. (4) APS does not deliver rated power or malfunctions. (5) Actuators malfunction.
Post-Test Exclusion	(1) Tests which fail due to causes external to subsystem *C*; e.g., other principal subsystem failures, test operator error, instrumentation or facility malfunction, provided that failure of subsystem *C* has not already occurred. (2) Failure of the subsystem due to the failure of an obsolete or experimental part so listed prior to the test on the declaration form. (3) Tests in which subsystem *C* did not fail but did not have the opportunity to satisfy the declared intention of the test.

Notes: It may be required that the engine contractor be held responsible for attaining a numerical reliability requirement. This implies that the numerical reliability requirement should not encompass: (1) equipment developed by another *associate contractor* which is tested in conjunction with the engine contractor's system; nor (2) interface attachments, the function of which is a joint responsibility between two or more associate contractors. When, however, equipment is furnished by *vendor* or *sub-contractor* to the engine contractor, and does not fall in category (2), the equipment is considered to be the engine contractor's responsibility with respect to meeting any numerical reliability requirements.

Declaration Policy

In order to establish the integrity of the data, it is necessary to determine the intentions of a test and by applying the ground rules for applicability establish the utility of the test results for reliability purpose. This is done by means of the declaration form which must be completed prior to each full-scale test as part of the test plan. It is then submitted to the program manager for approval. The form is self-explanatory, but it is appropriate to note that when exclusions are declared the reasons should be given, together with any substantiating information and/or references to failure reports, inspection reports, engineering change orders, etc. For example, if a part is provisionally excluded by reasons of obsolescence, the engineering change order number or the new part (drawing) number, etc., should be given on the form. (Fig. 1 is a sample declaration form.) The program manager might establish a limit to the number of provisional exclusions.

Reliability Reporting and Estimation

The reliability report form (Fig. 2 is a sample reliability report form) is completed as soon as possible after the test data are reduced. The data are then evaluated by the contractor's reliability group on the basis of the ground rules and the information contained on the declaration form. This is done for each principal subsystem after each full-scale test. A short description of the failure, or reason for exclusion when any principal subsystem is so classified, should appear in the remarks column. When a failure occurs which cannot be assigned to any particular subsystem, this fact should be noted as well. References to failure, and corrective action reports, etc., should also be given as necessary. The reliability report forms covering a particular calendar month can then be gathered in chronological order and summarized as illustrated in the example to follow. The best estimate and 95 per cent lower confidence limit of reliability can then be obtained by the methods given later and reported monthly.

Representative and Current Data

Because of the small number of tests available for evaluation, each principal subsystem must necessarily display a relatively high reliability, and few failures must, therefore, occur. However, at the beginning of the program several failures may occur, establishing such a high cumulative failure rate that, were there no provision for discarding data, an exorbitantly

DECLARATION FORM

A. Engine Serial No. ..... Engine Type ..... Test Stand .....

Test No. ..... Date of Test .....

B. Which of the following principal subsystems are applicable for reliability evaluation?

	Applicable	Not Applicable
(1) Propellant Charge-Ignition (*A*)		
(2) Case-Liner-Internal Insulation (*B*)		
(3) Thrust Vector (*C*)		

C. If any of the principal subsystems are not applicable state reasons.

.....

.....

.....

D. Which performance parameters should be excluded from reliability evaluation?

State reasons: .....

.....

.....

E. Which components or parts are declared experimental or obsolete (for provisional exclusion)?

.....

.....

.....

F. Comments: .....

.....

.....

G. Movable nozzle program: Number of cycles, angles of deflection, and period of operation.

.....

.....

Signatures and Approvals:

Program Management ..... Test Engineer .....

Date ..... Reliability .....

Fig. 1.

RELIABILITY REPORT FORM

Date of Test..... Test No..... Test Stand..... Test Report No.....

Engine Type..... Engine Serial No.....

Engine Test Data

Total Impulse, lbf-sec	.....	Ignition Delay, sec	.....
Average Thrust, lbf	.....	Ignition Peak Thrust, lbf	.....
Specific Impulse, lbf sec/lbm	.....	Useful Propellant Weight, lbm	.....
Max–Min (Thrust Time Curve), lbf	.....	Engine Weight, lbm	.....
Rate of Thrust Rise, lbf/sec	.....	Propellant Mass Fraction	.....
Rate of Thrust Decline, lbf/sec	.....	Firing Temperature, °F	.....
Action Time, sec	.....	Other Environments	.....
Average Chamber Pressure, psia	.....	.....	.....

Test Result Classification
(to be completed by reliability group)

Principal Subsystem	Exclusion		Failure		Success		Remarks
	Pre-test	Post-test	Perf Excl	Perf Not Excl	Perf Excl	No Perf Fail	
A							
B							
C							

Additional Remarks: .....

.....

.....

Responsible Engineer..... Department and Group..... Date.....

Fig. 2.

large number of successful tests would be needed before the earlier failure rate would be "absorbed". This is not feasible, and in a development program which expects to improve the product, it is not realistic to handle the data this way. If these failures were random failures or failures due to unassignable causes, then it would not be legitimate to discard the earlier data, as this failure pattern would be the manifestation of the inherent reliability and, as such, would indicate that the subsystem was not sufficiently reliable. However, the earlier failures are not generally random; i.e., they do have assignable causes, and in a development program are subject to analysis and corrective action.

This corrective action is called a reliability design change, which is defined as a modification to correct a previously observed failure and must be intended to appear on all subsequent engines. Data generated after the reliability design change will be regarded as homogeneous, and earlier data discarded as being no longer representative of the current design. Data produced after the reliability design change will be called "current data". The decision as to what constitutes a reliability design change is subject to the approval of the program manager and relates only to that particular subsystem for which corrective action has been taken.

All engine reliability estimates (as described

in the example below) will be based on current data only. Thus, when counting applicable tests the count should not be extended any further back than the last reliability design change for each principal subsystem.

Performance Reproducibility

Performance reproducibility is not incorporated into the reliability evaluation method described in this paper; i.e., it is not incorporated into the reliability estimates to be reported. However, generally, model specifications contain requirements relating to between-engines performance variation. Applicable data obtained from the reliability report form would be used to determine the probability of meeting these variability requirements. Further, if these pertinent performance parameters are suitably chosen; e.g., specific impulse, thrust, weight, etc., the probability of the missile falling within the CEP can be computed on a propulsion reliability basis.

Engine Reliability Estimation

*Technique of Estimation**

Engine reliability will be computed from principal subsystem test results. After each test, it will generally be possible to classify each of the principal subsystems as having succeeded, failed or been excluded. The exception to this classification can occur when there is a failure but it is not known which principal subsystem(s) failed. While it is desirable to be able to classify completely each test, an unassignable failure still represents a system failure and must be incorporated into any system reliability estimate. The general procedure is as follows: The number of applicable tests on each principal subsystem are determined. The minimum of these three numbers, N, gives the number of equivalent engine tests performed. To compute the reliability for the engine it is necessary to count the number of known failures which occurred in the last N tests of each subsystem. In addition, those failures which have not been assigned or attributed to any subsystem in particular are arbitrarily assigned to the individual principal subsystems. This is done for all possible arrangements consistent with the number of failures and tests appearing in the unassignable category. Thus, when the known failures are added to the arrangements, a set of all possible failure arrangements for a given number of equivalent engine tests is obtained. Any one failure arrangement can be written $(N; f_1, f_2, f_3)$, where f_1, f_2, f_3 are the number of failures of the separate subsystems; and $f_1 + f_2 + f_3$ is therefore the total number of assignable and unassignable failures during the period of N equivalent engine tests. The minimum reliability for 95 per cent confidence can then be obtained from the tables or graph for each $(N; f_1, f_2, f_3)$. Some of the numerical values of these reliabilities will be the same since the value of the reliability is independent of the order of f_1, f_2 and f_3.

The smallest of the reliabilities is taken as the minimum demonstrated reliability which could occur in a population represented by the sample. If later analysis permits unassignable failures to be allocated, the estimates will then be re-computed.

The best estimate of reliability for the same sample, is obtained by taking that combination of failures which gave the smallest of the reliabilities for the confidence estimate and substituting in $(N - f_1)\,(N - f_2)\,(N - f_3)/N^3$.

An Example of Estimation

(a) An example will illustrate the method of engine reliability estimation. The method of tabulating the results as shown in Table 4 will be found convenient. During each test all principal subsystems are physically present but may or may not be applicable for reliability evaluation according to the ground rules. The result of the test for each of subsystems A, B and C will be success (S), failure (F) or exclusion (which is indicated by a blank in the table). In addition, there may be unassignable failures. These are indicated under that combination of subsystems (AB, BC, AC or ABC) which contains the potential failing subsystem(s). The following Table 4 of hypothetical results, together with accompanying detailed explanation, will clarify the above discussion.

* The mathematical basis for the method of reliability estimation is based upon the work described in Ref. (1) and (2) and the tables in Ref. (3).

Table 4.

Test No. (chronological)	Reliability report No.	Date of test	Assignable results			Unassignable failures			
			A	*B*	*C*	*AB*	*BC*	*AC*	*ABC*
1			S						
2				F					
3			S		S*				
4						F			
5			S*	S*	F*				
6			S*				F*		
7			S*	S*	S*				
8									F*
9			S*	S*					

* See paragraph (b), below.

Subsystem	No. of Times tested	Assignable failures	Possible failures
A	8	0	2
B	7	1	3
C	5	1	2

Test No.	Explanation of tabulation
1	Subsystem *A* was applicable and a success. *B*, *C* were excluded.
2	Subsystem *B* was applicable and a failure. *A*, *C* were excluded.
3	Subsystems *A* and *C* were applicable and successes. *B* was excluded.
4	Subsystems *A* and *B* were both applicable but a failure occurred; post-test analysis failed to establish which of the two subsystems failed or whether it was both. *C* was excluded.
5	*A*, *B* were both successful; *C* a failure.
6	*A* was a success but an unassignable failure occurred in *B* and/or *C*.
7	All subsystems successful.
8	All subsystems applicable but (an) unassignable failure(s) occurred in one or more of them.
9	*A* and *B* were both successful. *C* was excluded.

(b) In the above example, the principal subsystem experiencing the fewest number of tests is *C* with 5; thus, a maximum of the equivalent of five complete engines have been tested. Counting back from test No. 9 for the last five tests of each principal subsystem, it is found that *A* was present and applicable in tests Nos. 9, 8, 7, 6, 5; subsystem *B* in tests Nos. 9, 8, 7, 6, 5 and subsystem *C* in tests Nos. 8, 7, 6, 5 and 3. Thus, only the results of the aforementioned tests for each subsystem should be utilized in obtaining the engine reliability estimate; i.e., only those results marked by an asterisk in the table. During the period of testing the five engines, *A* has no known failures and one possible failure in test No. 8. *B* has no known failures, two possible failures (one in test No. 6 and one in test No. 8). *C* has one known failure in test No. 5 and two possible failures in tests Nos. 6 and 8. All the possible arrangements of the unassignable failures are shown below.

Subsystem *A*	Subsystem *B*	Subsystem *C*	Total unassignable failures
0	0	2	2
0	1	1	2
0	2	0	2
1	0	1	2
1	1	0	2

The known failures: 0, 0, 1, respectively, are added to each of the above arrangements to give all possible variations, both known and unassignable; in the notation $(N; f_1, f_2, f_3)$ these become

(5; 0, 0, 3)
(5; 0, 1, 2)
(5; 0, 2, 1)
(5; 1, 0, 2)
(5; 1, 1, 1)

From Fig. 3 it is seen that the first combination gives a value of 7·6 per cent for minimum demonstrated reliability at the 95 per cent confidence level. The next three combinations give 16·3 per cent and the last combination gives 18·9 per cent.

The minimum of the possible reliabilities; i.e. 7·6 per cent with 95 per cent confidence is the reported reliability. In this example, based on hypothetical test results, the possible reliabilities vary considerably because of two facts. Firstly, there is a much greater proportion of unassignable failures than known failures and, therefore, the arrangements greatly affect the reliability estimates, and secondly, the sample size is very small. Generally, this will not be the case, so that the reliability estimates will be much closer together.

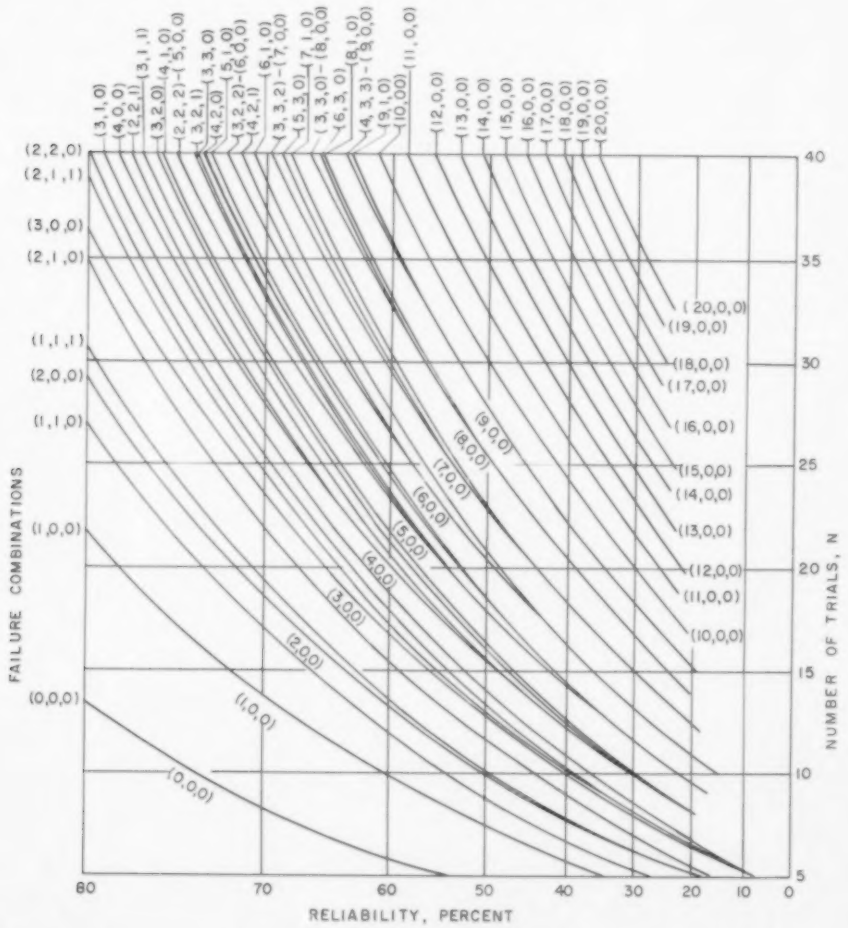


Fig. 3. 95 per cent lower confidence limit to true reliability for observed failure combinations of a three-subsystem serial system with $N(5 < N < 40)$ trials per subsystem. (All permutations of failure combinations (f_1, f_2, f_3) are equivalent.)

The best estimate of reliability, as obtained from the worst combination of failures (0, 0, 3), given by the formula is

$$(5 - 0)(5 - 0)(5 - 3)/5^3 = 2/5 = 40\%$$

Use of Fig. 3 and Table 5

Figure 3 gives the estimate of reliability at a 95 per cent confidence level for a given number of trials, $N = 5$–40, and a particular combination of failures (f_1, f_2, f_3). The figure covers values of $(f_1 + f_2 + f_3)$ from 0 to 20; however, not all failure combinations are plotted. Table 5 lists in descending order with respect to the reliability values, various failure combinations. Only those combinations marked with a single asterisk have been plotted. Thus, if it is necessary to obtain a reliability estimate for a combination not

Table 5. List of Failure Combinations in Order of Descending Reliability

Total No. of failures $(f_1 + f_2 + f_3)$	Combination of failures (f_1, f_2, f_3)	Total No. of failures $(f_1 + f_2 + f_3)$	Combination of failures (f_1, f_2, f_3)	Total No. of failures $(f_1 + f_2 + f_3)$	Combination of failures (f_1, f_2, f_3)
0	0 0 0*	8	3 3 2*	11†	10 0 0*
			4 2 2		
1	1 0 0*		4 3 1	12†	11 0 0*
			4 4 0		
2	1 1 0*		5 2 1	13†	12 0 0*
	2 0 0*		5 3 0*		
			6 1 1	14†	13 0 0*
3	1 1 1*		6 2 0		
	2 1 0*		7 1 0*	15†	14 0 0*
	3 0 0*		8 0 0*		
				16†	15 0 0*
4	2 1 1*	9	3 3 3*		
	2 2 0*		4 3 2	17†	16 0 0*
	3 1 0*		4 4 1		
	4 0 0*		5 2 2	18†	17 0 0*
			5 3 1		
5	2 2 1*		5 4 0	18†	18 0 0*
	3 1 1*		6 2 1		
	3 2 1*		6 3 0*	20†	19 0 0*
	4 1 0*		7 1 1		
	5 0 0*		7 2 0	21†	20 0 0*
			8 1 0*		
6	2 2 2*		9 0 0*		
	3 2 1*				
	3 3 0	10	4 3 3*		
	4 1 1		4 4 2		
	4 2 0*		5 3 2		
	5 1 0*		5 4 1		
	6 0 0*		5 5 0		
			6 2 2		
7	3 2 2*		6 3 1		
	3 3 1		6 4 0		
	4 2 1*		7 2 1		
	4 3 0		7 3 0		
	5 1 1		8 1 1		
	5 2 0		8 2 0		
	6 1 0*		9 1 0*		
	7 0 0*		10 0 0*		

Note: All permutations of f_1, f_2, f_3 are equivalent.

* Plotted in Fig. 3.

† For those values of $f_1 + f_2 + f_3 \geq 11$ the estimate is obtained from $[N; (f_1 + f_2 + f_3 - 1), 0, 0]$ unless two of f_1, f_2, f_3 (f_2, f_3, say) are zero, in which case the reliability is given by $(N; f_1, 0, 0)$.

plotted, e.g. (3, 3, 1), it can be done for any particular N by interpolating between the curves for (3, 2, 2) and (4, 2, 1). Ref. (3) allows reliability estimates to be obtained for a wider range of N and various confidence levels.

CONCLUSION

The paper has shown how the reliability of an engine can be estimated from engineering development tests without restricting the exploratory nature of the tests. These estimates can be obtained even though the configuration and/or hardware being evaluated has not reached its final design. The statistical technique used in the analysis of the data results in a more efficient confidence estimate and thereby supersedes the product rule for interval estimation.

REFERENCES

1. R. J. Buehler, *J. Amer. Stat. Assoc.* **52,** 482 (1957).
2. *Measurement of Over-All System Reliability Utilizing Results of Independent Subsystem Tests*, STL Report No. GM-TR-0165-00506 (Oct. 1958).
3. *Tables of Upper Confidence Limits on Failure Probability of 1, 2 and 3 Component Serial Systems* (2 Vols.), STL Report No. TR-59-0000-00756 (July 1959).

TECHNIQUES IN THE DEMONSTRATION OF RELIABILITY ATTAINMENT

EDWIN D. KARMIOL and JOHN S. YOUTCHEFF
Missile and Space Vehicle Department, General Electric Company

Abstract—This paper describes the analytical techniques which are being utilized in the demonstration of equipment reliability in a re-entry vehicle system development program. Analytical techniques have been established to provide equipment reliability predictions at periodic intervals as functional and environmental test data are compiled. A mathematical model has been developed which allows for a realistic and sound approach to the problem. The equations allow for the calculation of equipment reliability, at several confidence levels, in both the individual and combined operational environments. These expressions have been programmed on the IBM 704 computer to provide a completely mechanized reliability data handling and processing system. Equipment reliability status reports are issued periodically to provide management with a quantitative measure of equipment reliability, and offer the design engineer an opportunity for any needed redesign early in the program.

The primary objective of a reliability program is to attain and maintain equipment of a desired reliability level. This necessitates measuring the equipment parameters and predicting equipment operational reliability based on test data. It is essential that equipment reliability be measured as early in a development program as possible.

The test data provides a basis for confirming equipment suitability for system utilization and allows time for necessary corrective action to be taken on unreliable equipment. Although many techniques exist for evaluating reliability, the choice of methodology becomes limited when program time and cost factors are considered. It is, therefore, necessary to examine fully and integrate these control factors in the formulation of a practical reliability program. This paper describes specific reliability program measurement and prediction techniques established for use in the development of complex re-entry vehicle systems where time and cost factors are of high importance.

When considering the system operational reliability, three major areas should be investigated. The operational environments must first be ascertained, providing a valid indication of the environmental stresses in both the flight and preflight operations. The adequacy of the planned test activities must next be evaluated to assure that equipment environmental test requirements are at levels equal to or above operational stresses. Analytical methods must then be developed to make optimum use of all available test data in predicting the equipment survival probability over the required operational limits.

Re-entry vehicle system environmental considerations can be divided into two distinct operations. These are the preflight operation, covering the ground phases from system assembly to launch, and the flight operation, covering the flight trajectory from the point of launch to the target. The overall environmental requirement is a factor of the stresses to which the vehicle will be subjected in these two operational areas.

The flight operation is considered to be of prime importance because during flight the equipment is beyond the reach of maintenance activities while it is subjected to the highest operational stresses and functional demands. This flight operation can be subdivided into three major flight phases; the initial or launch phase, the midcourse phase and the terminal phase. For the type of re-entry vehicle we are considering here, these phases become the powered flight phase, the free flight phase and the re-entry phase. The powered flight phase can consist of one or more stages of active propulsion, terminating at engine cutoff which occurs after a specified time period when ballistic velocity is reached. The free flight phase

extends from engine cutoff to a point on the trajectory at which the re-entry vehicle re-enters the effective atmosphere. The re-entry phase extends from this point of atmospheric contact to the point of impact.

In each of these flight phases the environmental stresses can be analytically estimated for each stage of propulsion where the thermal stresses are increasing but are of short time duration. Equipment vibrational amplitudes and frequencies can also be estimated for each stage of powered flight where the vibrational stress is of the longest time period. Although the free flight phase accounts for the greatest

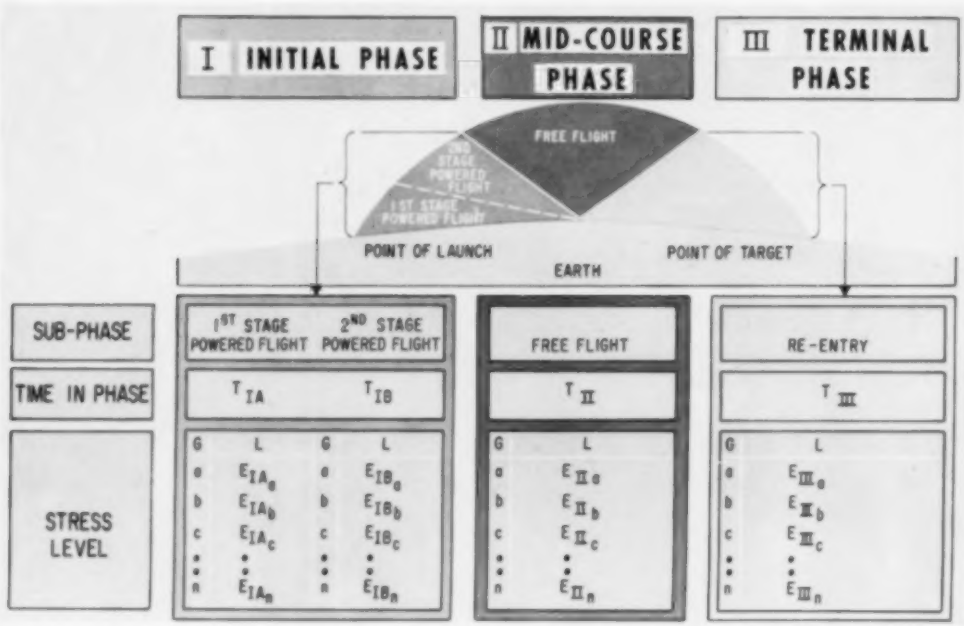


Fig. 1. Flight operation. G = group, L = level. For further notation see text.

component groups (Fig. 1). The notation used in Figs. 1 and 2 are as follows:

T	time in phase; flight operation
T′	time in phase; preflight operation
E	environmental stress level; flight operation
E′	environmental stress level; preflight operation
a, b, c, n	equipment group as subjected to flight environment
a′, b′, c′, n′	equipment group as subjected to preflight environment
e	test stress level
t	test duration.

For example, the equipment temperature environment can be estimated for each group at

magnitude of the flight time, operational conditions are such that several key environmental stresses such as vibration and acceleration are effectively zero.

While the re-entry vehicle is subjected to high environmental stresses during the re-entry phase, the time involved is of extremely short duration. It can be seen that environmental consideration can be directed to each of the phases of re-entry vehicle flight, where the stress levels and time durations can be analysed for the various equipment groupings.

The same approach can be taken in determining the environmental stresses in the preflight operation. This can be accomplished by considering each of the preflight operational phases which can be designated under the

functional phases of transportation, storage, and preflight checkout.

In analysing the environmental stresses of these various preflight operational phases, consideration must be given to packaging and other means that are used to protect the system and associated equipment from environments which are recognized as being detrimental. With this consideration, the ground environmental stresses to which the equipment may be subjected can be determined from each preflight operational phase (Fig. 2). The equipment environment can be estimated for each mode of transportation where the environmental stresses will depend upon the method of shipment. Although equipment storage times could be of extremely long duration, most environments can usually be controlled in this phase and no stresses of any consequence are likely to be encountered. In the preflight checkout phase, however, environmental stresses will be experienced during periods of captive flight where the engine and equipment are operating. As in the flight operation, equipment environmental levels can be estimated for the stresses to which the equipment may be subjected for various time periods within these preflight phases.

It is, then, possible to establish the stress levels for the various environments to which the system and associated equipment will be subjected in the several phases of the flight and preflight operation (Fig. 3). The time periods at the various stress levels and the equipment functional requirements within these time phases, can now be used with available test data in the prediction analysis.

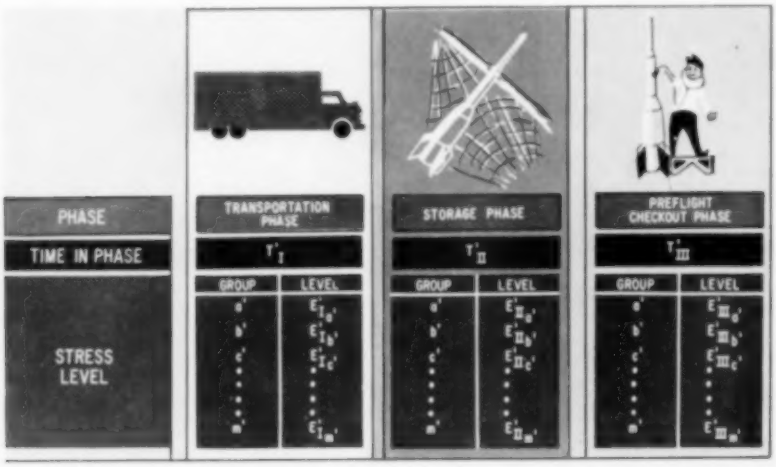


Fig. 2. Preflight operation. Notation see text. Primed letters are used to distinguish the preflight operational equipment stress levels from the flight operation equipment stress levels.

There are four major test activities in the system development program. These are: design qualification testing, equipment acceptance testing, field testing and flight testing. Each of these test activities provides valuable data relative to the equipment operation under specified environmental conditions. Within the overall test program, design qualification tests are of major importance in providing equipment environmental data. The equipment design qualification testing program provides for an examination of the stress levels to which the equipment may be subjected in operation, allowing the establishment of adequate environmental test limits (Fig. 4). These tests provide system operational data relative to the environmental stresses as specified. The equipment qualification program in conjunction with the field and flight tests,

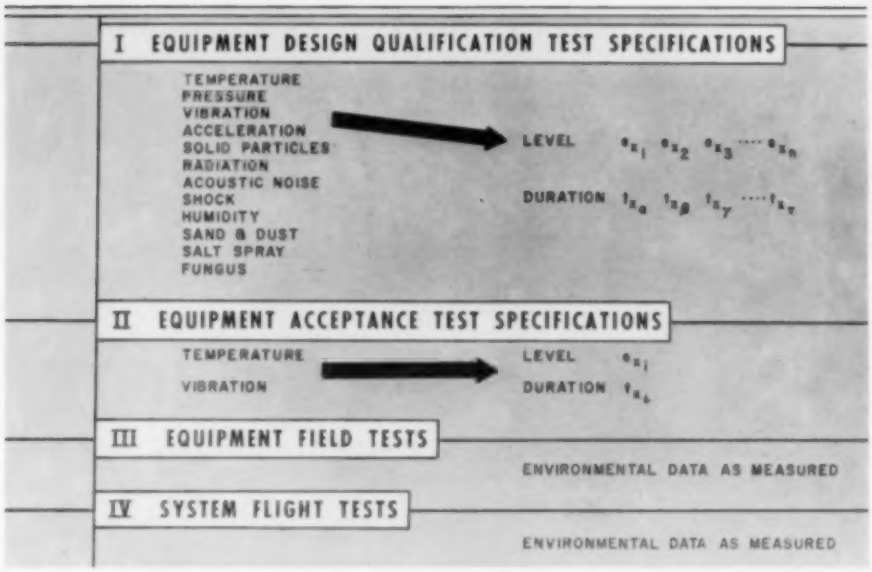


Fig. 3. Operational environments.

thus, provides the data necessary for evaluation of equipment reliability while engaged in proving design capability and performance. Equipment acceptance requirements include specific temperature and vibration testing to insure the detection of workmanship defects and incipient flaws in production items. This testing provides additional data regarding the behavior of the equipment under reduced thermal and vibration environments.

These environmental test data are used as inputs for the data analysis and mechanized reliability programs from which the probability of equipment survival under various operational

ENVIRONMENT	PREFLIGHT OPERATIONAL PHASE			FLIGHT OPERATIONAL PHASE			
	I	II	III	IA	IB	IIA	IIIA
TEMPERATURE			X	X	X	X	X
PRESSURE				X	X	X	X
VIBRATION	X		X	X	X		X
ACCELERATION				X	X		X
SOLID PARTICLES					X	X	
RADIATION					X	X	
ACOUSTIC NOISE			X	X	X		
SHOCK	X		X				
HUMIDITY	X		X				
SAND & DUST		X	X				
SALT SPRAY			X				
FUNGUS		X					

Fig. 4. Environmental test requirements.

stress levels can be determined (Fig. 5). All test data obtained throughout the development program are utilized to the fullest extent in the prediction of system survival in its functional environment.

In predicting system survival under the various environmental stresses, test data are accrued from all functional and environmental tests performed on both development and operational equipment. These test data can be recorded in a reliability data log containing provisions for equipment test time, test environment and test results (Fig. 6). This data log can provide a test history of all system equipment and can be used in the statistical analysis of equipment operational life.

The prediction analysis can include a testing plan, using equipment test data that is compiled throughout the development program. Basically, the equipment that is tested represents a sample, N, of one of more units subjected to phases of the operational environment, e_i, for a given test period, t_i. The number of failures, X, if any, is noted and the results are used in estimating the probability of survival of this type of equipment in the operational environments for the required periods. Here the probability of equipment survival can be predicted in the various operational environments. This is done by computing the component operational failure rate for the environmental stresses considered.[1] The maximum likelihood estimate is

$$\lambda_i = \left[1 + \frac{1}{2} N_i\right]^{-1} \frac{\Sigma X_i}{\Sigma t_i \alpha_i} \tag{1}$$

where α is a correction factor relating the test environment to the operational environment.

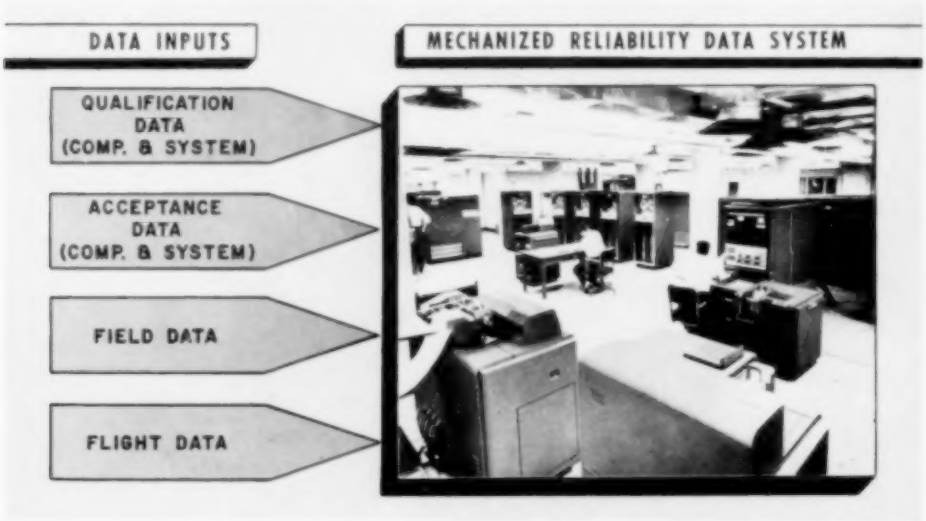


Fig. 5. Mechanized reliability program.

Here, the operational failure rate upper bound can be computed to any desired confidence as determined by the choice of the normal deviate, K.

$$\lambda_{i_u} = \frac{2\lambda_i + \dfrac{K^2}{\Sigma t_i \alpha_i} + \sqrt{\left[4\lambda_i \dfrac{K^2}{\Sigma t_i \alpha_i} + \dfrac{K^4}{(\Sigma t_i \alpha_i)^2}\right]}}{2} \tag{2}$$

or if, $\lambda_i = 0$, $\lambda_{i_u} = \dfrac{K^2}{\Sigma t_i \alpha_i}$ (3)

The system operational failure rate can be found as

$$\lambda_s = \Sigma_i \lambda_i \tag{4}$$

The upper bound on this failure rate can

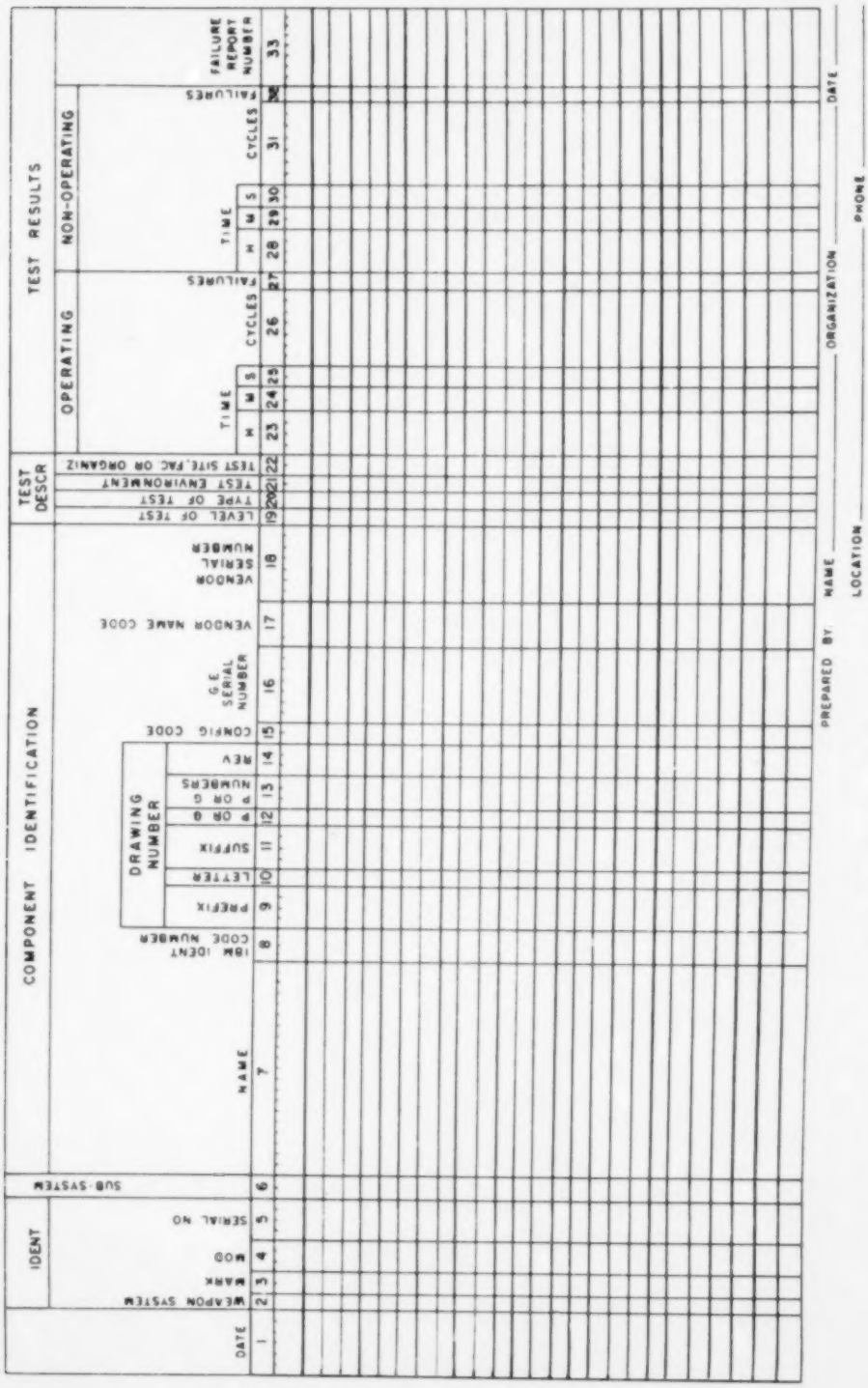


Fig. 6. Reliability test data accumulation and recording.

again be computed to the desired confidence level:

$$\lambda_{s_u} = \frac{2\lambda_s + K^2C + \sqrt{(4\lambda_s K^2C + K^4C^2)}}{2} \quad (5)$$

where

$$C = \frac{\Sigma \frac{\lambda_i}{t_i a_i}}{\lambda_s}$$

Reliability indices can be computed from the predicted failure rate for both the components and system. The reliability best estimate can be expressed as

$$R = \exp - \lambda \quad (6)$$

The reliability lower bound can also be found as

$$R_l = \exp - \lambda u \quad (7)$$

It is also possible to determine the amount and type of testing required to insure a specified survival probability with a given confidence. Tables based on a sequential life test plan can also be utilized for obtaining the equipment survival probability values for various test times and test results, sequentially compiled during equipment testing (Figs. 7 and 8).[2] These tables provide a means of keeping track of equipment reliability in a given operational environment as the test results are accumulated during the various test phases of the system development program.

These prediction methods require that the equipment failures in the area of interest are subject to an exponential chance–failure curve.

Reliability	Unit minutes of failure free testing required per operational minute			
	Confidence			
	0·95	0·90	0·80	0·50
0·9999	29960	23030	16090	6930
0·9998	14980	11515	8045	3464
0·9997	9986	7677	5363	2310
0·9996	7490	5758	4022	1732
0·9995	5992	4606	3218	1386
0·9994	4993	3838	2682	1155
0·9993	4280	3290	2299	990
0·9992	3745	3879	2011	866
0·9991	3329	2559	1788	770
0·9990	2996	2303	1609	693
0·9980	1498	1152	805	347
0·9970	999	768	536	231
0·9960	749	576	402	173
0·9950	599	461	322	139
0·9940	499	384	268	116
0·9930	428	329	230	99
0·9920	375	288	201	87
0·9910	333	256	179	77
0·9900	300	230	161	69
0·9800	150	115	81	35
0·9700	100	77	54	23
0·9600	75	58	40	17
0·9500	59	45	32	14
0·9400	48	37	26	11
0·9300	41	32	22	9
0·9200	36	28	19	8
0·9100	32	25	17	7
0·9000	29	22	15	7
0·8000	14	10	7	3
0·7000	8	6	5	2
0·6000	6	5	3	1
0·5000	4	3	2	1

Fig. 7. Test time requirements.

Unit minutes of testing per operational min	95% Confidence						90% Confidence					
	0 failures	1 failure	2 failures	3 failures	4 failures	5 failures	0 failures	1 failure	2 failures	3 failures	4 failures	5 failures
1	0·05	0·01	0·002	0·0004	0·0001	0·00005	0·10	0 02	0·005	0·001	0·0003	0·0001
2	0·22	0·09	0·04	0·02	0·01	0·005	0·32	0·14	0·07	0·04	0·02	0·009
3	0·37	0·20	0·12	0·08	0·05	0·03	0·47	0·28	0·17	0·11	0·07	0·05
4	0·47	0·30	0·21	0·15	0·10	0·07	0·56	0·37	0·25	0·19	0·13	0·09
5	0·55	0·38	0·28	0·21	0·16	0·12	0·63	0·46	0·35	0·26	0·20	0·15
6	0·61	0·45	0·35	0·28	0·22	0·17	0·68	0·52	0·42	0·33	0·27	0·21
7	0·65	0·50	0·41	0·33	0·27	0·22	0·72	0·57	0·47	0·38	0·32	0·26
8	0·68	0·54	0·45	0·38	0·31	0·26	0·75	0·61	0·51	0·43	0·36	0·30
9	0·72	0·59	0·50	0·43	0·36	0·31	0·77	0·64	0·55	0·47	0·41	0·35
10	0·74	0·62	0·53	0·46	0·40	0·35	0·79	0·68	0·59	0·51	0·45	0·39
20	0·86	0·79	0·73	0·68	0·63	0·59	0·89	0·82	0·77	0·72	0·67	0·62
30	0·900	0·85	0·81	0·77	0·73	0·70	0·924	0·88	0·84	0·80	0·77	0·73
40	0·925	0·89	0·85	0·83	0·79	0·77	0·942	0·90	0·88	0·84	0·82	0 79
50	0·940	0·90	0·88	0·86	0·83	0·81	0·954	0·922	0·90	0·88	0·85	0·83
60	0·950	0·920	0·90	0·88	0·86	0·84	0·962	0·935	0·913	0·90	0·88	0·86
70	0·957	0·931	0·910	0·90	0·88	0·86	0·967	0·944	0·924	0·910	0·89	0·87
80	0·962	0·939	0·920	0·90	0·89	0·88	0·971	0·951	0·933	0·916	0·90	0·89
90	0·967	0·947	0·931	0·915	0·90	0·89	0·974	0·956	0·940	0·925	0·910	0·90
100	0·970	0·952	0·947	0·923	0·908	0·900	0·977	0·961	0·947	0·933	0·920	
200	0·985	0·976	0·968	0·961	0·954	0·948	0·988	0·980	0·974	0·967	0·960	0·953
300	0·990	0·984	0·979	0·974	0·969	0·965	0·9924	0·987	0·982	0·980	0·974	0·969
400	0·9925	0·988	0·984	0·981	0·977	0·974	0·9942	0·9901	0·987	0·983	0·980	0·976
500	0·9940	0·9904	0·987	0·985	0·982	0·979	0·9954	0·9922	0·989	0·987	0·984	0·981
600	0·9950	0·9920	0·990	0·987	0·985	0·982	0·9962	0·9935	0·9913	0·989	0·987	0·984
700	0·9957	0·9931	0·9910	0·989	0·987	0·985	0·9967	0·9944	0·9924	0·9904	0·988	0·986
800	0·9962	0·9939	0·9920	0·990	0·988	0·987	0·9971	0·9951	0·9933	0·9916	0·990	0·988
900	0·9967	0·9957	0·9931	0·9915	0·989	0·988	0·9974	0·9956	0·9940	0·9925	0·9910	0·989
1000	0·9970	0·9952	0·9947	0·9923	0·9980	0·9900	0·9977	0·9961	0·9947	0·9933	0·9920	0·9906
2000	0·9988	0·9976	0·9968	0·9961	0·9954	0·9948	0·9988	0·9980	0·9974	0·9967	0·9960	0·9953
3000	0·9990	0·9984	0·9979	0·9974	0·9969	0·9965	0·9992	0·9987	0·9982	0·9980	0·9974	0·9969
4000	0·9993	0·9988	0·9984	0·9981	0·9977	0·9974	0·9994	0·9990	0·9987	0·9983	0·9980	0·9976
5000	0·9994	0·9990	0·9987	0·9985	0·9982	0·9979	0·9995	0·9992	0·9989	0·9987	0·9984	0·9981
6000	0·9995	0·9992	0·9990	0·9987	0·9985	0·9985	0·9982	0·9996	0·9991	0·9989	0·9987	0·9984
7000	0·9996	0·9993	0·9991	0·9989	0·9987	0·9985	0·9997	0·9994	0·9992	0·9990	0·9988	0·9986
8000	0·9996	0·9994	0·9992	0·9990	0·9988	0·9987	0·9997	0·9995	0·9993	0·9992	0·9990	0·9988
9000	0·9997	0·9995	0·9993	0·9991	0·9989	0·9988	0·9997	0·9996	0·9993	0·9993	0·9991	0·9989
10000	0·99970	0·99952	0·99947	0·99923	0·99908	0·99900	0·99977	0·99961	0·99947	0·99933	0·99920	0·99906
20000	0·99985	0·99926	0·99968	0·99961	0·99954	0·99948	0·99988	0·99980	0·99974	0·99967	0·99960	0·99953
30000	0·99990	0·99984	0·99979	0·99974	0·99969	0·99965	0·99992	0·99987	0·99982	0·99980	0·99974	0·99969
40000	0·99993	0·99988	0·99984	0·99981	0·99977	0·99974	0·99994	0·99990	0·99987	0·99983	0·99980	0·99976
50000	0·99994	0·99990	0·99987	0·99985	0·99982	0·99979	0·99995	0·99992	0·99989	0·99987	0·99984	0·99981
60000	0·99995	0·99992	0·99990	0·99987	0·99985	0·99982	0·99996	0·99994	0·99991	0·99989	0·99987	0·99984
70000	0·99996	0·99993	0·99991	0·99989	0·99987	0·99985	0·99997	0·99994	0·99992	0·99990	0·99988	0·99986
80000	0·99996	0·99994	0·99992	0·99990	0·99988	0·99987	0·99997	0·99995	0·99993	0·99992	0·99990	0·99988
90000	0·99997	0·99995	0·99993	0·99991	0·99899	0·99988	0·99997	0·99996	0·99994	0·99993	0·99991	
100000	0·99997	0·99995	0·99994	0·99992	0·99991	0·99990	0·99998	0·99996	0·99995	0·99993	0·99992	0·99991

Fig. 8. Equipment survival probability.

This can be verified or modified as test data are accrued on the various system components. In obtaining the test data, the system equipment on which the unit test time is accrued must be from the same population as the operational equipment. The test environment for a given test period must also be as severe or more severe than the flight environment for the operational time period considered. Care must be taken that the equipment wear-out portion of the life curve is well beyond the total test and flight time in the required operational environments.

A mechanized program can be established to provide for the efficient handling and processing of all data necessary for the adequate analysis of equipment reliability. A mechanized program permits large quantities of data to be accumulated, analysed and processed to provide the necessary equipment reliability indices and other vital reliability data at periodic intervals throughout the system development program.

The data are accrued in accordance with the requirements of the reliability test data accumulation and failure reporting specifications. All test data inputs are recorded as mentioned, in the reliability data log, and equipment failures are analysed in accordance with the requirements of the reliability failure analysis specification. The reliability data log provides a complete equipment operational history from initial component acceptance through the system flight test operation. It also provides an equipment test history which can be used in the statistical analysis of the operational life of the equipment.

The test data that are collected include equipment identification, applicable test specification, test environment, operating time and associated failures, if any. These data are recorded on IBM cards and processed on the IBM 704 computer in the following arrangements:

Sort I provides the test time that each serialized component has been subjected to each environment, and the quantity of components tested and failed.

Sort II provides the total test time that a particular component type has been subjected to each environment and the failures which occurred therein.

Sort III provides the total test time and failures logged against each serialized component.

Sort IV provides the total test time and failures logged against each component type.

The data are then automatically processed and evaluated relative to the equipment operational environments and required time therein. This is done through programmed steps on the IBM 704 computer, using the equations developed for reliability prediction purposes.

Equipment reliability status reports are automatically generated and issued periodically. These reports present a complete listing of equipment failure rates and reliability indices for the Atlas and Thor re-entry vehicle components, subsystems and systems, for each anticipated environmental condition and overall operation (Figs. 9 and 10). These output reports show equipment reliability status and reliability growth which can be graphically presented over the development cycle (Fig. 11). They also point out areas where concentrated effort will give the greatest gains in terms of improved reliability. These reports have provided management with a quantitative measure of equipment reliability status while affording the design engineer an opportunity for any needed redesign early in the development program.

The re-entry vehicle reliability prediction program provides:

1. Reliability information as to the component survival probability in each of the operational environments.
2. A determination of which components require immediate attention, prior to their assembly in the system.
3. Data necessary for establishing the equipment survival probability in the combined operational environment.

(SUPPLEMENT)

IBM CODE NO	COMPONENT DRAWING NO.	TEMPERATURE			VIBRATION			ACCELERATION A			ACCELERATION B			LIFE			OPERATION	
		$\bar{\lambda}$	λ	λ_u	$\bar{\lambda}$	λ	λ_u	$\bar{\lambda}$	λ	λ_u	$\bar{\lambda}$	λ	λ_u	$\bar{\lambda}$	λ	λ_u	λ	λ_u
	SYSTEM																	
A01A	0000A0001P001			.00000			.00001			.00001			.00001	.00000	.00000	.00002	.00000	.00002
A01A	0000A0001P001			.00000			.00001			.00001			.00001	.00000	.00000	.00002	.00000	.00002
A01A	0000A0001P001			.00000			.00001			.00001			.00001	.00000	.00000	.00002	.00000	.00002
B01A	0194A0302P001													.00033	.00066	.00074	.00066	.00074
B01A	0194A0302P001													.00033	.00066	.00074	.00066	.00074
B01A	0194A0302P001													.00033	.00066	.00074	.00066	.00074
B01A	0194A0302P001													.00033	.00066	.00074	.00066	.00074
C01A	0194A0303P001															.00000		.00011
C01A	0194A0303P001															.00000		.00011
D01A	0199A0544P001			.00001			.00125			.00001			.00001			.00000		.03045
D01A	0199A0544P001			.00001			.00125			.00001			.00001			.00000		.03045
D01A	0199A0544P001			.00001			.00125			.00001			.00001			.00000		.03045
E01A	0179A0695P001			.00009			.01082									.00001		.09120
F01A	0859A0806P001			.00000			.00000							.00005	.00010	.00014	.00010	.00014
F01A	0859A0806P001			.00000			.00000							.00005	.00010	.00014	.00010	.00014
F01A	0859A0806P001			.00000			.00000							.00005	.00010	.00014	.00010	.00014
F01A	0859A0806P001			.00000			.00000							.00005	.00010	.00014	.00010	.00014
F01A	0859A0806P001			.00000			.00000							.00005	.00010	.00014	.00010	.00014
G01A	0932A0367P001			.00001			.00014							.00008	.00259	.00577	.00259	.00577
H01A	0932A0348P001			.00001			.00028							.00013	.00410	.00843	.00410	.00843
I01A	0932A0349P001			.00001			.00035			.00001			.00001	.00002	.00062	.00280	.00062	.00280
J01A	0859A0810P001			.00018			.00089			.00001			.00001	.00012	.00368	.01116	.00368	.01116
K01A	0859A0811P001			.00008			.00044							.00006	.00170	.00767	.00170	.00767
L01A	0219A0812P001	.00052	.00052	.00100	.00027	.00160	.00718							.00007	.00263	.00373	.00474	.00851
L01A	0219A0812P001	.00052	.00052	.00100	.00027	.00160	.00718							.00007	.00263	.00373	.00474	.00851
	SYSTEM														LAMBDA		.02531	

$\bar{\lambda}$ = Failure rate per minute (in specified environment)

λ = Failure rate best estimate per operation

λ_u = Failure rate upper bound (.95 Confidence) per operation

DATE SAMPLE PAGE SAMPLE

Fig. 9. Equipment reliability status report (supplement).

IBM CODE NO.	COMPONENT NAME	TEMPERATURE						VIBRATION						ACCEL. A				ACCEL. B				LIFE						REL.	
		QUAL.		ACCEP		REL.		QUAL.		ACCEP		REL.		QUAL.		REL.		QUAL.		REL.		QUAL.		ACCEP		REL.			
		T	F	T	F	50	95	T	F	T	F	50	95	T	F	50	95	T	F	50	95	T	F	T	F	50	95	50	95
	SYSTEM																												
A01A	Component A			209		1	.999			283		1	.999	2		1	.999	2		1	.999	12		451	1	.999	.999	.999	.999
A01A	Component A			209		1	.999			283		1	.999	2		1	.999	2		1	.999	12		451	1	.999	.999	.999	.999
A01A	Component A			209		1	.999			283		1	.999	2		1	.999	2		1	.999	12		451	1	.999	.999	.999	.999
B01A	Component B																					4		112	2	.999	.996	.999	.999
B01A	Component B																					4		112	2	.999	.999	.999	.999
B01A	Component B																					4		112	2	.999	.999	.999	.999
B01A	Component B																					4		112	2	.999	.999	.999	.999
C01A	Component C																					2		45	1	.999	1	.999	
C01A	Component C																					2		45	1	.999	1	.999	
D01A	Component D			72		1	.999			72		1	.998	3		1	.999	3		1	.999	9		174	1	.999	1	.978	
D01A	Component D			72		1	.999			72		1	.998	3		1	.999	3		1	.999	9		174	1	.999	1	.978	
D01A	Component D			72		1	.999			72		1	.998	3		1	.999	3		1	.999	9		174	1	.995	1	.978	
E01A	Component E			25		1	.999			25		1	.981									4		50	1	.999	1	.912	
F01A	Component F			599		1	.999			517		1	.999									14		849	28	.999	.999	.999	.999
F01A	Component F			599		1	.999			517		1	.999									14		849	28	.999	.999	.999	.999
F01A	Component F			599		1	.999			517		1	.999									14		849	28	.999	.999	.999	.999
F01A	Component F			599		1	.999			517		1	.999									14		849	28	.999	.999	.999	.999
F01A	Component F			599		1	.999			517		1	.999									14		849	28	.999	.999	.999	.999
G01A	Component G			148		1	.999			143		1	.999									4		205	4	.997	.994	.997	.994
H01A	Component H			148		1	.999			149		1	.999									4		227	5	.995	.991	.995	.991
I01A	Component I			139		1	.999			139		1	.999	2		1	.999			.999		5		212	1	.999	.997	.999	.997
J01A	Component J			42		1	.999			42		1	.999	2		1	.999	2		1	.999			97	2	.996	.998	.996	.988
K01A	Component K			40		1	.999			41		1	.999											80	1	.998	.992	.998	.992
L01A	Component L			287	6	.999	.999			150	1	.998	.992									4		431	22	.997	.996	.995	.991
L01A	Component L			287	6	.999	.999			150	1	.998	.992									4		431	22	.997	.996	.995	.991
	SYSTEM																										R	.974	

LEGEND
T = Number of components subjected to test (in specified environment)
F = Number of components failing in test (in specified environment)
R_{50} = Reliability best estimate per operation
R_{95} = Reliability lower bound (.95 Confidence) per operation

DATE SAMPLE PAGE SAMPLE

Fig. 10. Equipment reliability status report.

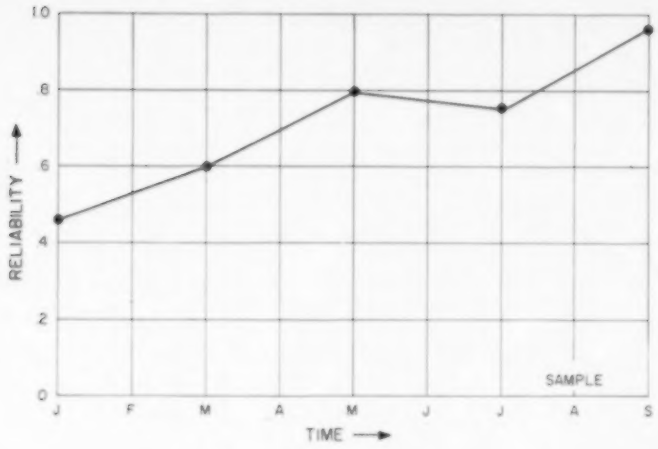


Fig. 11. System reliability.

4. Data for adequate equipment provisioning at the earliest possible date.

Reliability prediction program activities have been designed to provide the above listed objectives, as necessary, throughout the system development. The value of the described reliability prediction techniques has been established on ballistic missile and space vehicle systems, and other vital defense projects where high operational survival is mandatory.

REFERENCES

1. D. Rubinstein, *Statistical Methods on System Reliability*, General Electric Company, General Engineering Laboratory (June 24, 1958).
2. E. L. Fritz and J. S. Youtcheff, *Sequential Life Testing in Systems Development*, General Electric Company MSVD (May 1, 1957).

ANALYSIS OF RANDOM FAILURES

G. P. ANDERSON and L. E. PETERSON
Remington Rand Univac, Division of Sperry Rand Corporation

Abstract—The term "random failure" has been used much in recent literature to justify those few failures of electronic parts which could not readily be explained. The goal seems to have been the elimination of causes of failures until only random failures remain. The inference is that nothing can be done about these.

It is the purpose of this paper to show how analysis of random failures has been used in product improvement at Remington Rand Univac. Described are some of the techniques employed in the analysis of computer sealed chassis and component part failures. The paper also points out that these same techniques can be used in evaluating components proposed for use in reliable equipment where insufficient time is available to conduct life evaluations.

INTRODUCTION

Early in 1955 RRU began work on the development of a large-scale, general purpose, transistorized, digital computer to be used as part of the radio guidance system for the Air Force ICBM Titan. A unique aspect of the contract covering the work was the fact that it placed specific requirements on the manufacturer regarding the desired level of reliability to be obtained. This level of reliability requires, effectively, that no more than six out of every 1000 missile firings shall fail to hit the target as a result of errors or failures on the part of the computer.

In view of the complexity of modern large-scale computers, and the large number of components involved, it was obvious that more than ordinary care would be required in circuit design and component selection to meet the reliability requirement. It was also apparent that all failures, (especially those occurring during the developmental and early manufacturing periods), would have to be studied carefully for information which would lead to product improvement and increased reliability. In this respect recent literature makes frequent reference to the term "random failure" which seems to cover those few failures of electronic parts for which there is no ready explanation. It was determined early in the developmental program at RRU that no failure could be regarded simply as "routine", "to be expected now and then" or "random".

This paper will concern itself with that portion of RRU's total reliability effort which is devoted to failure analysis. It will show that the analysis of failures which could be regarded as "random" can lead to product improvement and increased reliability.

System and Component Description

The Athena* Computer for the Titan Ground Guidance System is a large-scale digital machine which uses approximately 30,000 diodes, 10,000 transistors and 35,000 resistors. Fig. 1 shows the Athena computer complete with its power supply and peripheral equipment. The rollout of one computer bay is shown in the open position in Figs. 2 and 3 so that the hermetically sealed chassis may be seen in place. These chassis contain the logical elements of the computer, and the details of their construction are shown in Figs. 4–6. Each chassis contains two printed circuit boards made of fiberglass-epoxy laminate with copper foil on one side. These chassis are made with extreme care. Their assembly is a "clean room" operation with temperature and humidity control, white smocks and gloves for all operators, and strict process control. All components are hand-mounted on the circuit boards, and are dip-soldered on a selective soldering machine which permits masking so as to expose only those areas where solder is desired. Finally, the chassis cover is secured by induction soldering, the unit is baked and

* Trademark of Sperry Rand Corporation.

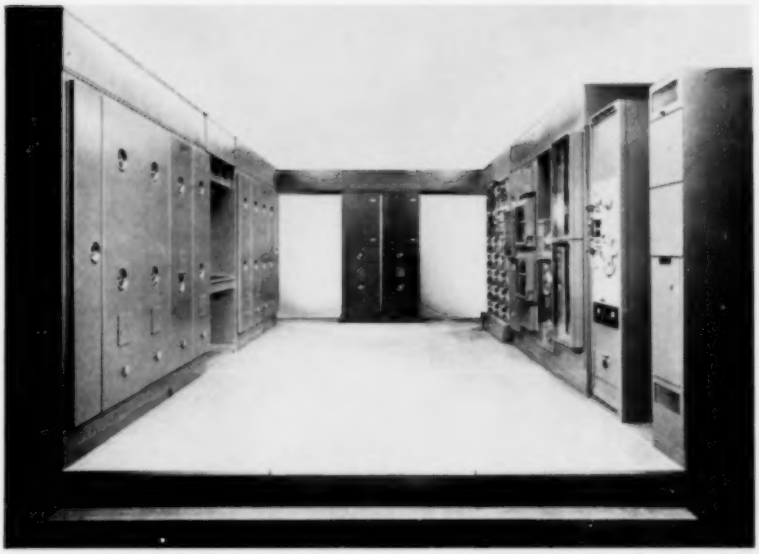


Fig. 1.

Fig. 2.

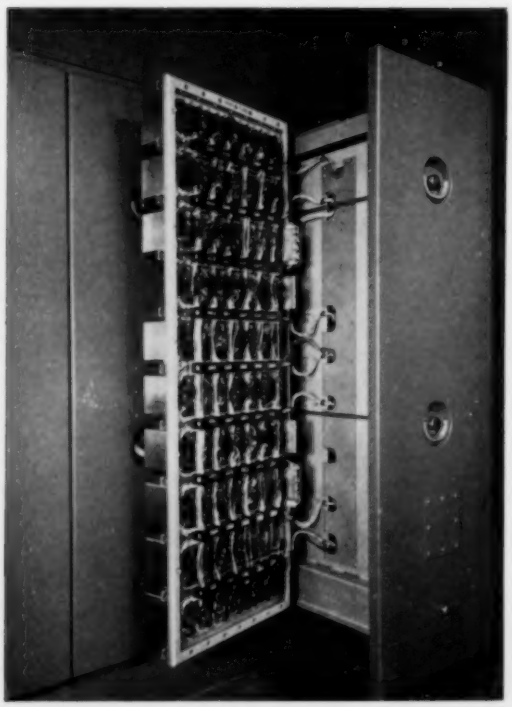


Fig. 3.

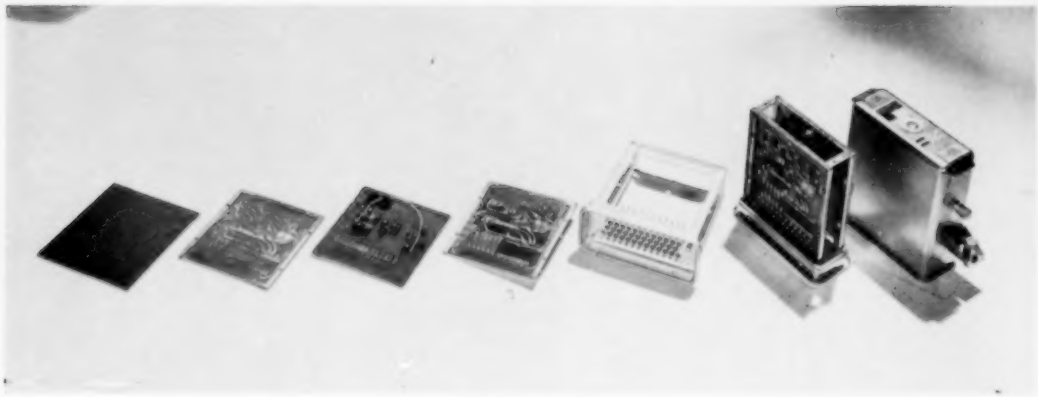


Fig. 4.

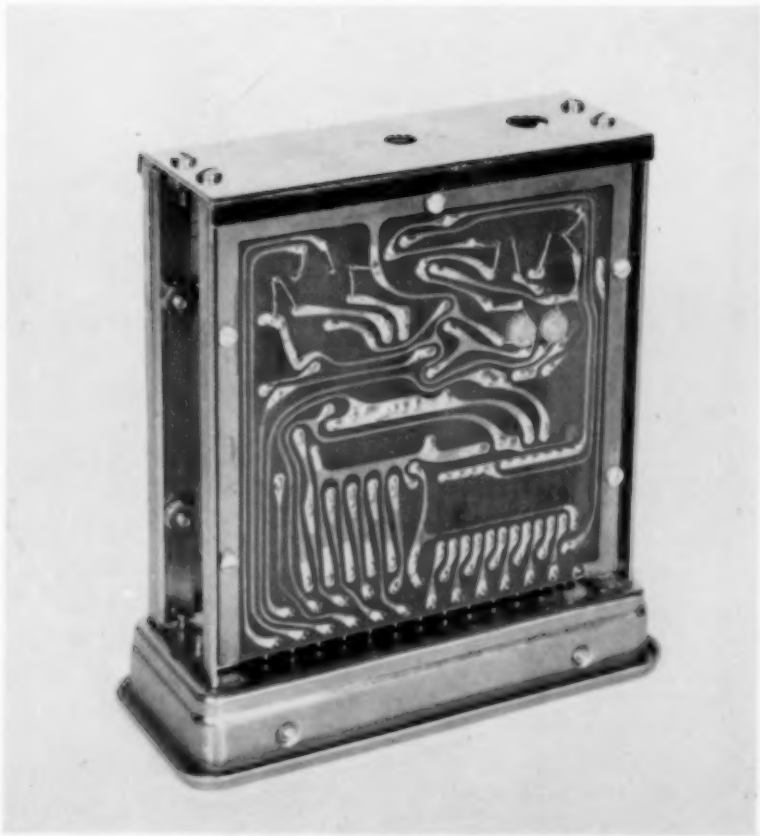


Fig. 5.

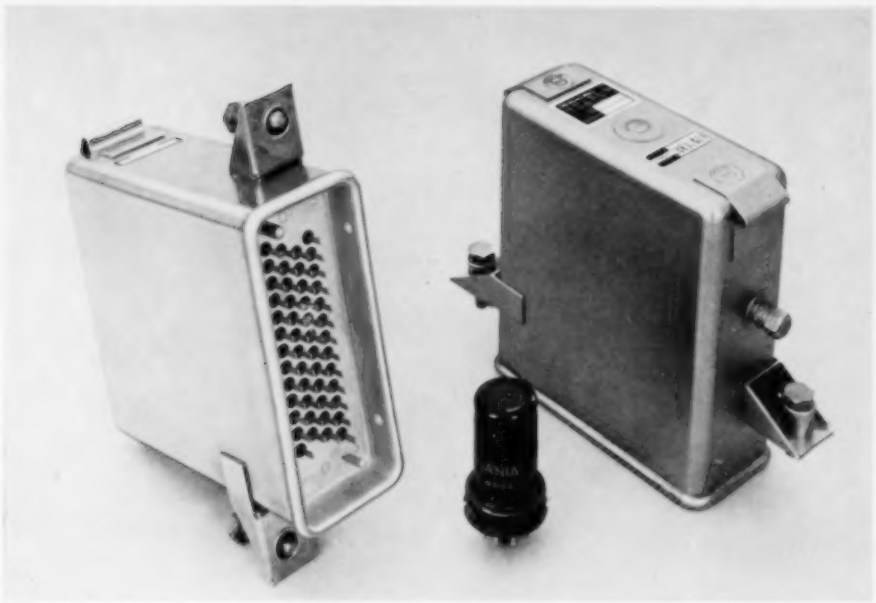


Fig. 6.

evacuated, and then pressurized with a mixture of dry nitrogen and helium. The entire computer is designed and assembled with reliability as the basic consideration. The failure records for the four Athena computers in the field show that none of the failures experienced to date in approximately 13,200 hr of machine operation has followed a pattern. All might be considered as random.

Failure Examples

A good example of the dividends that can accrue as a result of the careful analysis of an apparently prosaic failure occurred during the development period of Athena, Serial II. A sealed chassis used as a magnetic drum head amplifier displayed intermittent operation. A rather complex analytical tool known as a chassis tester was used first to locate the region of failure within the chassis. Additional measurements made externally at the chassis connector indicated a shorted transistor. Because of good external evidence the chassis cover was carefully unsealed, and by direct measurements at the suspected component the diagnosis of shorted transistor was "confirmed". At this point the removal of the transistor would be normally justified, and its replacement, in this case, with a new transistor would have corrected the

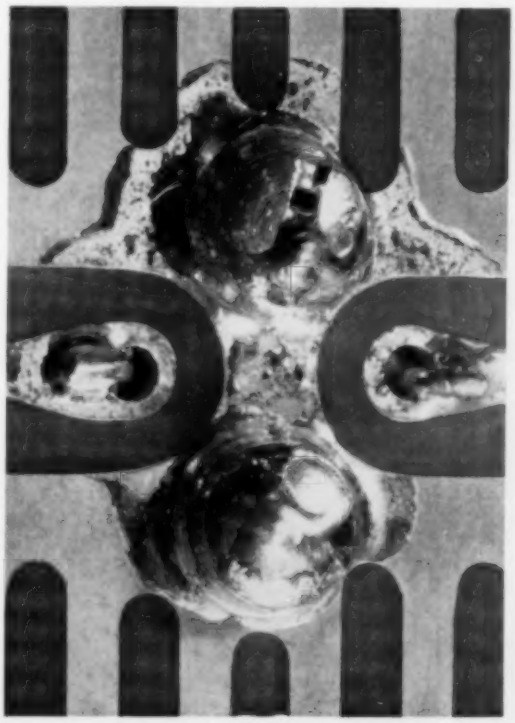


Fig. 7.

observed condition. Another random transistor failure would have been logged.

It was decided, however, to look a little deeper. In examining the printed circuit board the analyst noted that the holes through which the collector and base leads of the transistor pass were not completely covered with solder (see Fig. 7). Because the case of the suspected transistor was also the emitter electrode in this instance, it occurred to the analyst that the trouble could be associated with those open holes. With a microscope it was determined that "solder garbage" from the dip-soldering process had been forced up into these holes, but it could not be determined whether this material had bridged the glass seal-insulator of the transistor to produce the observed short.

Because the evidence of failure here could be easily disturbed by the very dissection process necessary to pin-point the failure, some means had to be devised to "fix" this evidence during dissection. After some research the "means" was found in one of the thermo-setting polyester resins. This plastic will set at a temperature which is low enough to prevent damage to semiconductor devices. By means of a pindrill a small hole was drilled through the printed circuit board near the partially soldered transistor lead hole. The exact point selected is shown in the cross section drawing of Fig. 8. It can be seen that the transistor has a depression, or shallow well, near its glass insulator-seal which facilitated matters. A hypodermic syringe was then used to flow cold plastic slowly through the pin hole into the suspected region. Another pin hole at the opposite side and the incompletely soldered lead hole were used for relief, and plastic was fed until it oozed through these holes. Using the same technique the transistor itself was also encapsulated.

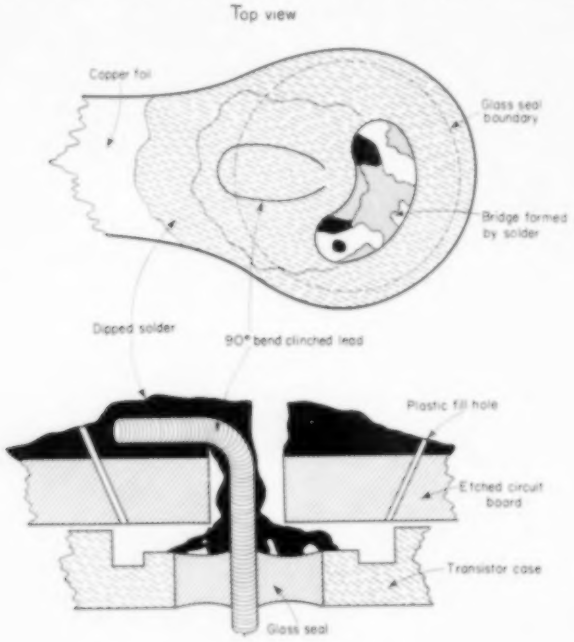


Fig. 8. Solder bridge failure at transistor lead hole.

When setting of the plastic was complete, dissection began. The results are shown in Figs. 9 and 10. The solder bridge was actually the cause of failure, and when it was removed the encapsulated transistor displayed normal operation.

As a result of this investigation several changes were made in the design of the printed circuit

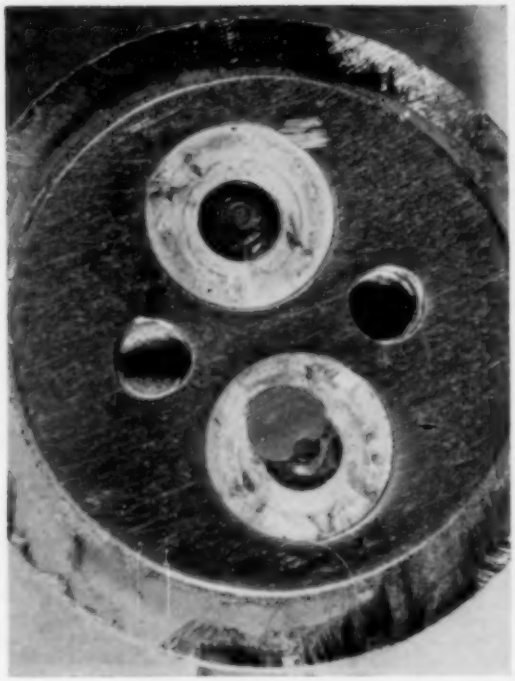


Fig. 9.

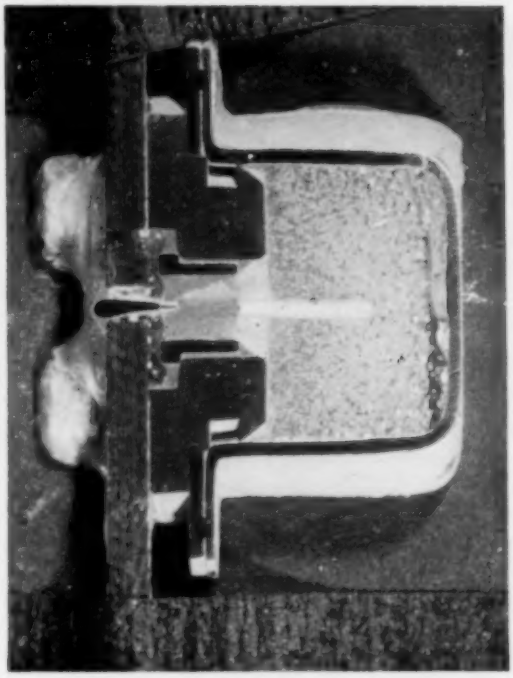


Fig. 10.

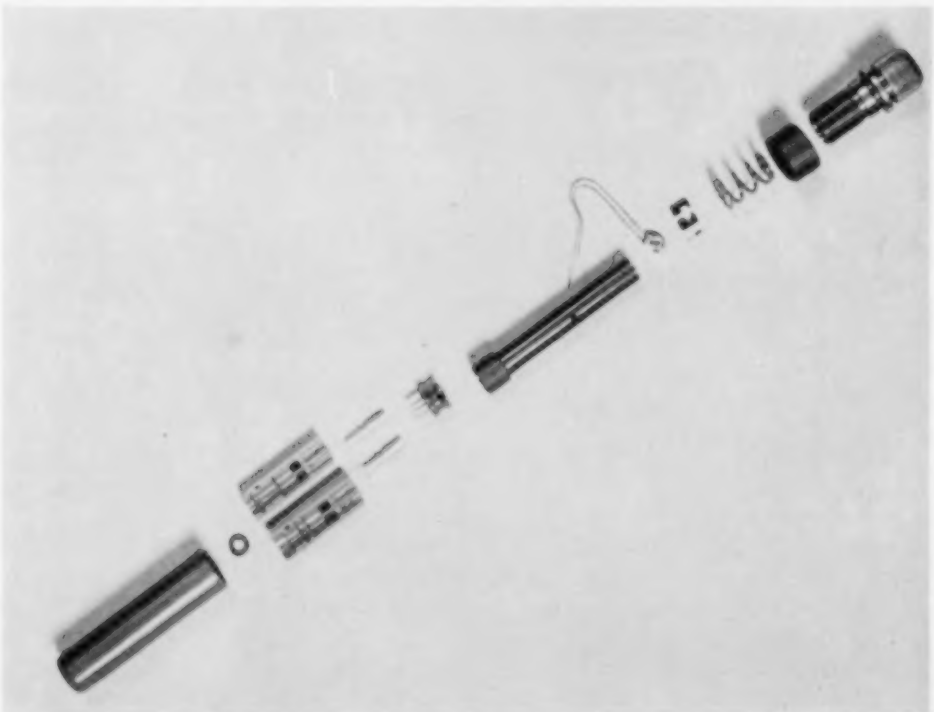


Fig. 11.

boards involved. Obviously the size of the transistor lead holes had to be reduced. They had been made large originally because of rather large manufacturing tolerances on the spacing of these leads. A dissection of the transistor involved, and several others like it removed from

Fig. 12.

printed circuit boards, showed cracked hermetic seals (Fig. 10). These were traceable to the bending over and clinching of the leads to the printed circuit boards since the leads were made of steel (with copper and gold overlays). The transistor leads are now dressed short and are, of course, no longer bent.

Another example of interest involved Athena, Serial IV, which produced occasional random errors. From the console information at the time of an error, the trouble was thought to be due to an intermittent chassis failure. The errors were very infrequent, and considerable time was required to collect data. Selective diagnostic procedures subsequently narrowed the difficulty down to the drum memory itself, and finally to

Fig. 13.

two suspicious magnetic drum heads. It was shown that substituting spare heads corrected the observed errors. Considering the number of these heads used in a system, these two could have been regarded as the sort of random failure you have to expect now and then, replaced by good heads and discarded.

Figure 11 shows the construction of these heads. The ferrite core is made in halves and fitted into the coil bobbin. An "air gap" is maintained by a thin silver shim. By means of a bent spring and potting compound the entire assembly is held firmly in the anodized aluminum shell.

It was first observed that both heads displayed a resistance leakage path to the shell (i.e. ground). Because this suggested internal damage (perhaps due to assembly), and because of the possibility of destroying subtle evidence in a blind attempt at dissection, the suspicious heads were X-rayed. The results of these X-rays are shown in Figs. 12 and 13. A possible cause of failure was seen in the position of the silver foil which appeared to be in contact with the coil. If this was true it could account for the observed shunt path to ground. In this case also, it could be reasoned that the actual area of contact with the shim would be small indeed and quite easily disturbed by dissection. Using the X-rays as a guide it was

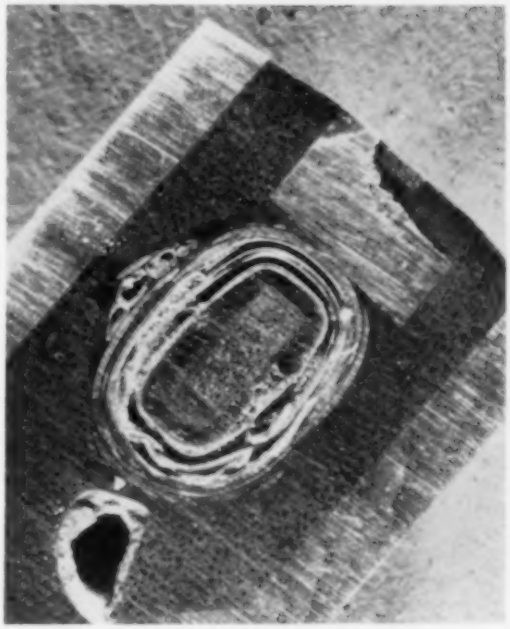


Fig. 14.

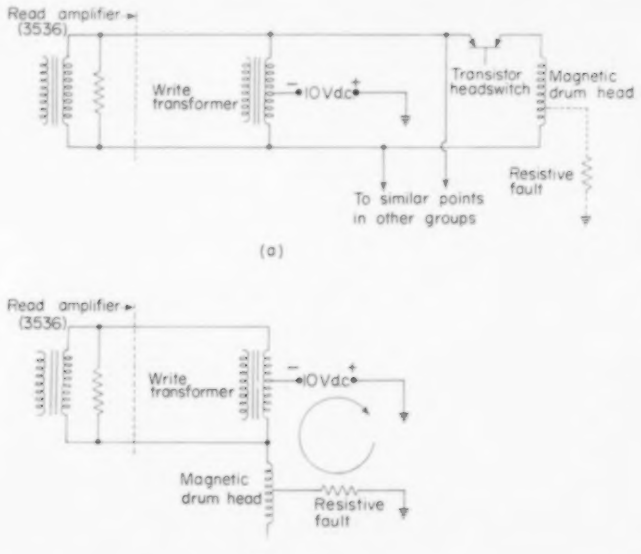


Fig. 15. Equivalent circuit analysis showing how defective drum head could cause errors. (a) Partial read–write schematic diagram. (b) Simplified equivalent circuit. A resistive short to ground from the coil in the magnetic drum head provides a path for ground noise and power supply noise through the −10 V supply to the read amplifier input. This causes bits to be read when the headswitch is open.

decided to lap into the suspected area in a plane parallel to the plane of the shim. By proper selection of abrasives and the application of care to the lapping process, it was possible to reveal the precise point of failure. From the results of the dissection as shown in Fig. 14, it seems likely that assemblers were using the coil as a backstop or positioning device for the silver shim. The introduction of the potting compound under pressure may well have helped to establish electrical contact between coil and shim. Simplified equivalent circuits of Fig. 15 show how the fault permitted power supply and ground noise pulses to be picked up in the low level read circuit.

In view of the above, certain design and assembly changes were made to prevent recurrence of these failures, and all such heads in the field measured to eliminate other potential failures. As in the case of the transistor mounting problem, needed design changes would not have been made had it not been decided to analyse these random failures carefully.

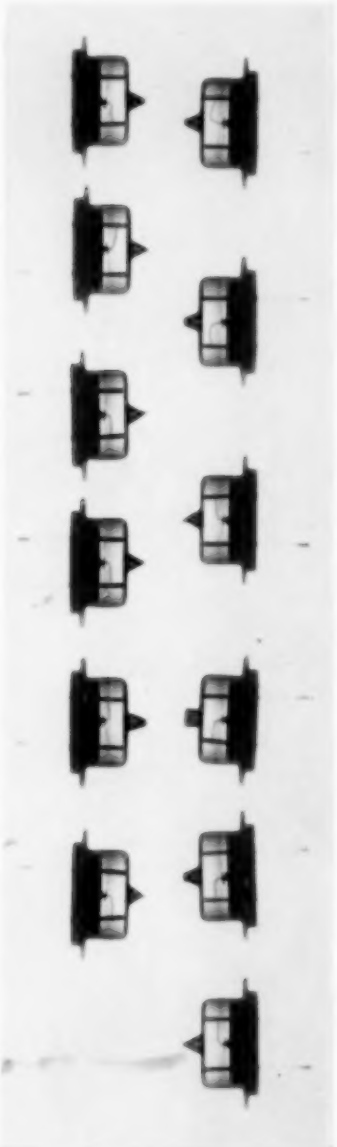


Fig. 16.

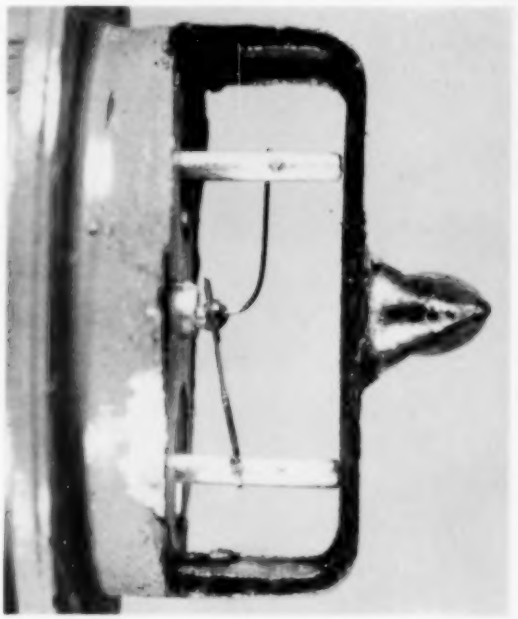


Fig. 17.

Occasionally an investigation of an "ordinary" failure will result in the discovery of a component manufacturing problem. Some rather surprising defects have been discovered which would have seriously affected the reliability of equipment using them. A recent example of this involved a power transistor used in the Athena Computer. The manufacturer decided to change the flat cap, or cover, of the transistor to the pinch-off style and still retain the same overall height for the unit. Effectively this meant that the available

height inside the unit was reduced by approximately $\frac{1}{8}$ in. An alert inspector noticed that one or two of the newly redesigned transistors which failed early due to internal short circuits were actually intermittent and also pressure sensitive. That is, they could be made to short by slight pressures applied to the case of the transistor. A random sample of new units was therefore examined by X-ray with the results shown in Fig. 16. It was apparent that whereas the top cap had been redesigned, the old header, with its long internal electrodes, was still in use (see Fig. 17). These electrodes varied slightly in

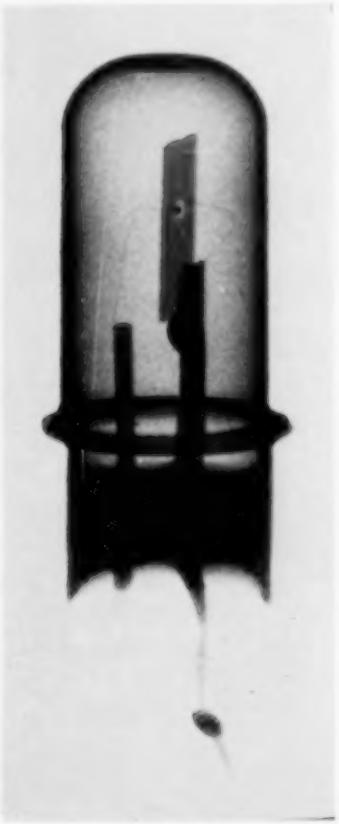


Fig. 18.

length from unit to unit, but quite frequently were long enough to produce a 3-way short of the transistor. Still others were perilously close; just waiting to join the ranks of future intermittent failures.

The manufacturer was contacted immediately and told of our findings. The defect had not yet been discovered at the factory, but immediate steps were taken to recall all deliveries of the new units. An interesting sidelight here is the fact that the manufacturer had observed a rather high rejection rate during production due to internal short circuits. It seems likely, however, that he chose to regard these as random failures.

In another instance, a failure which would have affected a guidance program was traced to an open emitter in a surface barrier transistor. This was an interesting failure since these units usually fail as a result of a collector–emitter short. Occasionally when they were used in an unprotected test circuit they might short first, and then burn open. But this was not possible in the computer circuit, and no other component in the failed circuit showed evidence of malfunction. It was decided to have a look inside the failed transistor.

Surface barrier transistors are quite small, and not too easy to dissect properly. Again, the X-ray was used to record the situation inside before dissection was begun. Fig. 18 shows that the emitter had pulled free from the indium dot on the base. The unit was carefully opened and examined under a microscope. The "kink" or bend in the emitter lead which appeared in the X-ray was now plainly evident (Fig. 19) and apparently responsible for the failure. These collector and emitter leads are only 1 mil in diameter, but they are quite stiff. Thus, the bend in this lead "spring loaded" it, and under tension, it slowly worked its way through the soft indium until it snapped free. The path it traveled in the indium could also be seen under the microscope. As a final check, using a home-made micromanipulator, the loose lead was carefully placed back into the indium from which it had broken free, and the unit again displayed transistor action. In this case the analysis had vindicated our circuit design.

A rather useful by-product of failure analysis methods in use at RRU is the final example. Recently, it became necessary to make a choice of a silicon power rectifier from two apparently equal components. The time factor made it

impossible to conduct comparative life testing. It was decided that in addition to short-term electrical evaluations random samples of these components would undergo dissection studies. Some of these were failures from prototype test units, some were failures from accelerated life tests, and some were new units.

In addition to somewhat poorer electrical behavior on the part of one unit it was shown that it had some rather poor mechanical design

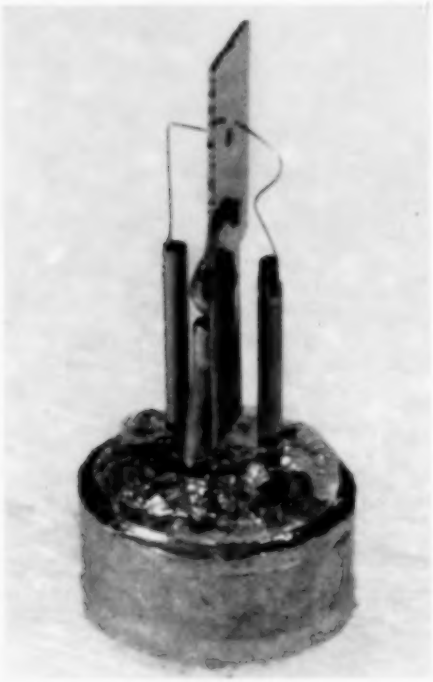


Fig. 19.

features. A seven-conductor stranded cable constitutes the flexible anode lead, and each end of this lead was probably dipped in what appears to be a high lead content solder. One end was then crimped into a connector lug, and the other end crimped into the terminal top of the rectifier case. Several of these terminations were electrically and mechanically unsound. Furthermore, the center conductor of the anode cable was not well-soldered at its terminations in most cases (Figs. 20 and 21). In those units which were encapsulated in plastic and dissected, some were found with small metallic "balls" inside

Fig. 20.

the unit (Fig. 22). In another the heat of normal operation was beginning to warp the silicon

Fig. 21.

Fig. 22.

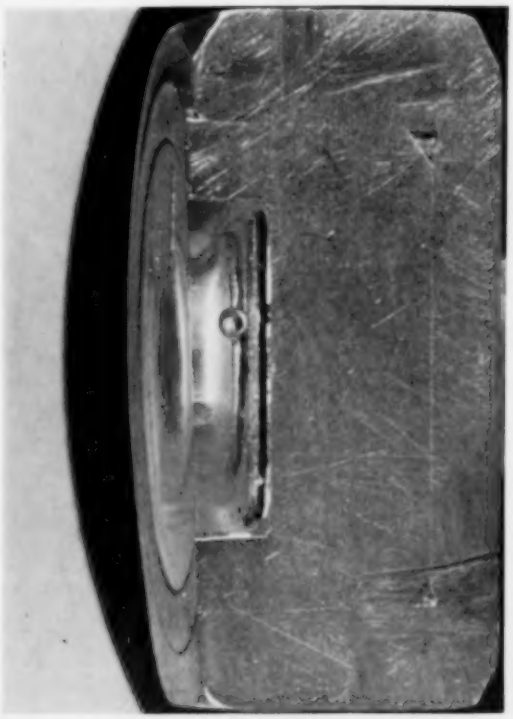


Fig. 23.

crystal away from its contact with the copper case (cathode connection, see Fig. 23). With this much additional evidence it was not difficult to decide in favor of the other unit.

CONCLUSIONS

Much space is given over in the literature these days to the increasing importance of reliability in complex weapons systems. There are frequent demonstrations of the fact that the failure of some small or inexpensive component can result in a failed mission costing millions of dollars. It is certain that there are many approaches to the problem of achieving this necessary increased reliability. At Remington Rand Univac we have attempted to achieve a "reliability attitude" which requires, among other things, close attention to the subtle details of reliability problems. One of these subtle details is the analysis of random failures.

It has been the purpose of this paper to point out advantages which have resulted from the careful analysis of failures at Remington Rand Univac. In several selected examples the dangers of disregarding the "infrequent", "ordinary" or "one of a kind" failure were pointed out. It was shown that these failures can sometimes hide needed design changes. It was also shown that even in the case of components with good reliability records, trouble can occur, and that failure analysis can prevent these components from finding their way into reliable equipment. By another example it was demonstrated that the methods of failure analysis can be used to aid in the total evaluation of components being considered for high reliability systems.

The four Athena Computers have now accumulated operational experience in excess of 13,200 machine hr with individual mean times to failure ranging from 203 hr to 533 hr. Careful attention to minor details in the selection of components, the design of the circuitry, the packaging of components and the manufacturing process is responsible in large measure for these records. It is also apparent at Remington Rand Univac that no intense reliability effort is complete without careful analysis of random failures.

RADIO COMMUNICATION BY MEANS OF SATELLITES

C. C. CUTLER
Bell Telephone Laboratories, Inc.

Abstract—This paper considers many of the problems and possibilities of long distance communication via satellite radio relays. An attempt is made to indicate plausible system proportions based on economic factors.

It appears that, where the situation is not complicated by tactical operations, very large ground terminals are justified in order to minimize satellite complexity and weight. Tactical systems, involving ground terminal mobility, almost certainly will be limited to narrow band service.

Passive satellites or very simple active transponders at an optimum satellite altitude of 1000–2000 miles could be erected within the present state of the art to provide broad band transoceanic service. With the development of more reliable components, rocket propulsion and control, one may justify using more sophisticated repeaters at 22,300 miles (24-hr orbit) with an active system of orientation control to give highly directional radiation toward the ground.

Many of the problems of mutual visibility, including optimizing orbits and numbers of satellites, are considered. It is found, for instance, that there is an optimum orbit altitude of around 1200 miles for passive (balloon) satellite repeaters, and of around 2000 miles for non-oriented or spinning active repeaters. Optimizing the proportions of active repeaters for minimum weight indicates that low powers and high gain antennas are called for.

The use of a large doubly curved passive reflector appears to be competitive with an active transponder in the case of the 24-hr satellite repeater.

INTRODUCTION

It seems almost certain that Earth satellites will be used as relays in transoceanic radio communication in the near future. It is not at all clear, however, what sort of system is most advantageous, nor how soon one could be implemented. The purpose of this paper is to present material helpful in considering this matter.

Many new technological developments make satellite relays look more practical than they did a few years ago. Not only do we have large propulsion systems and a demonstrated ability to put satellites into orbit, but we now have masers with a thousandfold increase in receiver sensitivity, an advanced knowledge of cosmic radiation levels, giant antennas, more efficient modulation systems and sources of very large r.f. power. Altogether, these factors take satellite communication out of the category of science fiction and make it imminently realizable.

Many of the factors involved in the design of a system are of a general nature. Problems of propagation, modulation systems, noise and power efficiency are pertinent to all systems and can be considered in general terms. Other problems, such as susceptibility to jamming, secrecy and portability apply particularly to military and tactical applications. Compatibility with existing systems, reliability, orbit stability and long life are particularly important to commercial (telephone system) use.

Some of the problems are novel. Thus it is unusual for a communication system to involve repeater acquisition and tracking. The repeaters are remarkably inaccessible for maintenance and are in an entirely new environment. The transmission path is almost entirely outside the atmosphere and traverses the ionosphere. The delay in transmission is much greater than usually experienced, and signal polarization is variable with satellite position. Also, there is the question of remote orientation and control of the repeaters. None of these problems appear beyond reach of today's technology, but each will require considerable attention before useful systems can be built.

A factor of great concern is the state of the art. Expediency would require a quite different system than would long range economics. The present high cost of satellite erection makes it seem daring, at present, to think in terms of a

few light satellites in crude orbits of a few hundred miles altitude. Systems based upon such satellites are clearly within the state of the art, however, and would have immediate usefulness. Ultimately one will want complex satellites at an altitude of 22,300 miles in extremely accurate equatorial 24-hr orbits where they will stay at fixed positions with respect to the surface of the Earth. Such a system seems far beyond our present abilities, but this story would not be complete without a consideration of the ultimate system, and it is not too early to begin research directed toward such an objective.

Both active and passive repeaters have been proposed.[1, 2] Pierce and Kompfner[3] have described in some detail a system using passive spherical satellites as reflectors. Such a system would require large tracking antennas and high power on the ground. Other passive reflectors, both planar and doubly curved, would be much smaller for the same signal transmission, but require various degrees of orientation control. Several possible repeater configurations are shown in Figs. 1 and 2 and are described in the next section.

Even a comparatively small and light active repeater could send to Earth a far stronger signal than could be obtained from a passive repeater, and would require only a modest power on the ground. However, the reliability or life of an active repeater, including both power supply and radio equipment is at present unevaluated. Further, as compared with passive repeaters, active repeaters have limited bandwidth and linearity, and hence have limited channel capacity and flexibility. It would, for instance, probably be undesirable to use a single amplifier for several simultaneous but independent signals because of the cross-talk problem.

For large volume point to point communication and for some military applications, large terminal installations are necessary to guarantee freedom from interference from other radio sources. Such system requirements, including reliability and long life, make passive repeaters look most promising. For military systems requiring relatively little traffic to mobile locations, terminal simplicity and portability are paramount, and an active satellite repeater is required.

After considering various repeater configurations we will discuss the state of the art which is pertinent to the design of satellite relay systems, including methods of modulation, antenna problems, propagation and interference, and system requirements. Finally we will present several system examples based upon these principles. Some of this material is also presented, with somewhat different emphasis in *Advances in Space Science* published in 1959 by Academic Press.

COMPARISON OF SATELLITE REPEATERS

A large variety of repeater configurations have been suggested from time to time. Several are shown in Figs. 1 and 2. There are:

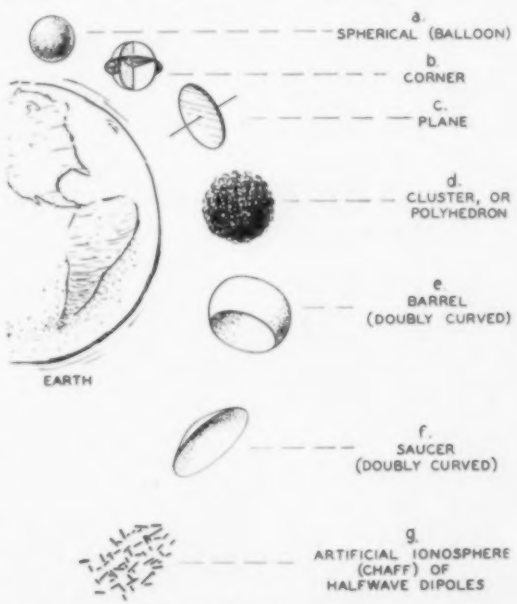


Fig. 1. Several forms of passive reflector which may be useful for satellite radio relay.

(1) A spherical (balloon) reflector (Fig. 1 (a)). This has been described at length by Pierce and Kompfner.[3] To a close approximation a sphere re-radiates energy isotropically and thus gives an attenuation between ground terminals of

$$\frac{P_T}{P_R} = \frac{16\lambda^2 p_1^2 p_2^2}{A_1 A_2 \eta^2 D^2} \qquad (1)$$

where λ is the wavelength, p_1 and p_2 the range to the respective terminals, A the antenna areas, η the antenna efficiency and D the sphere diameter. The balloon has the great advantage of simplicity and possibly very long life. The large value of space attenuation makes it imperative that very large ground installations be used, and relatively low altitudes. It is definitely not practical for the 24-hr synchronized orbit.

(2) A corner reflector (Fig. 1 (b)). If a satellite corner reflector is large enough to reflect a useful signal, it directs the signal straight back to its source with enough directivity to cover only a small ground area and therefore is of strictly limited usefulness.

The maximum permissible half power beam width of a corner reflector is about

$$\theta_c = 2r/p \tag{2}$$

where r is the distance from ground source to receiver, and p is the minimum satellite range.

This gives a directivity gain considered as a transmitter or receiver of*

$$G_c \cong \eta \frac{4\pi A_c}{\lambda^2} = \frac{7{\cdot}5}{\theta_c^2} = 1{\cdot}9 \left(\frac{p}{r}\right)^2 \tag{3}$$

where A_c is the effective area of the corner reflector. Using this relation with the free space propagation equation, and remembering that the receiver is necessarily a half beam width removed from the beam axis, the minimum path attenuation using a corner reflector is approximately

$$\left(\frac{P_T}{P_R}\right)_{\min} = \left(4{\cdot}7 \frac{\pi r^2}{A\eta}\right)^2 \tag{4}$$

Interestingly enough, this is independent of altitude, or satellite range. The reflector size is not independent of altitude, however, as can be seen from equation (3). The attenuation using a corner reflector is definitely prohibitive for a terminal separation longer than a few hundred miles.

The average attenuation may be reduced n-fold by using a cluster of n corner reflectors. Such a system gives a scintillating return (having large variations in amplitude) and necessarily involves some dispersion in time. Therefore it does not look desirable for any ordinary wide bandwidth communication purpose.

We will not consider the corner reflector further in this paper.

(3) An oriented plane reflector (Fig. 1 (c)). A plane reflector returns the energy in a sharply directed beam which must be pointed by accurately orienting the reflector. The great advantage of the much smaller path attenuation provided by use of the plane reflector must be balanced against the disadvantage of requiring a source of power and orientation control apparatus.

The intercepted energy from a disc is returned with a directivity gain of approximately

$$\pi^2 D^2/\lambda^2 \tag{5}$$

over an isotropic radiator and is contained in a beam width of

$$\theta = \frac{\lambda}{D}. \tag{6}$$

The path attenuation is then

$$\frac{P_T}{P_R} = \frac{16\lambda^4 p_1^2 p_2^2}{\pi^2 A_1 A_2 \eta_t^2 D^4} = \frac{16 p_1^2 p_2^2}{\pi^2 A_1 A_2 \eta_t^2} \cdot \theta^4 . \tag{7}$$

To be useful, such a reflector must be oriented with an angular accuracy smaller than $\theta/2$, and for low path loss the area of ground coverage is small. At high frequencies this seriously limits the practical reflector size.

A large number of plane reflectors may be grouped, as in a polyhedron (Fig. 1 (d)). This is a glinting type of reflector which returns much more peak power than an equal sized sphere, but with no more average power, and with a signal that is not likely to be useful for wide bandwidth communications.

(4) An oriented curved reflector. At high frequencies it is advantageous to use a curved passive reflector, so that the reflector can be made large enough to intercept a greater amount of the incident energy and reflect a broader beam of radiation than a plane reflector. There are two configurations of particular interest. The barrel-shaped reflector (Fig. 1 (e)) has the advantage that it can be spin stabilized on an axis

* This relationship between gain and beam width is empirical.

perpendicular to the plane of the orbit and requires no further attention. The saucer-shaped reflector (Fig. 1 (f)) on the other hand must be continuously oriented but makes much more efficient use of the incident energy. Either may consist of a specially shaped conductively coated balloon. The curved oriented reflector gives much less attenuation than either a comparable sized sphere or a planar reflector having comparable directivity.

Assuming that the reflector is large enough so that geometric optics can be applied, the beam width will be

$$\theta = 2\, w\rho \geqq \frac{r}{p} \tag{8}$$

where w, the reflector width is measured along the curvature. In case of the barrel, the curvature ρ is different in the two directions and wide fan beam is reflected with a width θ in the plane defined by the axis and the incident ray. The path attenuation* is approximately

$$\frac{P_T}{P_R} = \frac{16\lambda^2 p_1^2 p_2^2}{A_1 A_2 \eta^2 D^2} \cdot \frac{\theta}{2\pi} \cdot \tag{9}$$

For the saucer it is

$$\frac{P_T}{P_R} = \frac{16\lambda^2 p^4}{A^2 \eta^2 D^2} \cdot \frac{\theta^2}{4} \cdot \tag{10}$$

Unlike the planar reflector, the reflected beam width is nearly independent of frequency, and the path attenuation is proportional to the square of frequency instead of the fourth power. It is possible to reduce the path attenuation independently of reflector directivity by using larger reflectors and thus to optimize a compromise between reflector size and directivity. Because of edge diffraction, however, it is not practical to use a simple section of a sphere. Rather, the edge must be serrated or more sharply curved in order to reduce diffraction effects.

The curved reflector can be made more rigid, has less strict tolerance requirements than a planar reflector and can give reasonable path attenuations. Thus it appears to be competitive with other types of repeaters.

(5) A distribution of a multitude of tiny reflectors forming an artificial ionosphere (Fig. 1 (g)). This is tempting because it does not involve satellite acquisition and tracking. It requires an amount of propulsion to establish in orbit which is comparable to other satellites giving the same path attenuation. It is unacceptable for many applications however because the dispersion in time and frequency of returned energy due to the finite thickness of the layer and the range of Doppler shift caused by relative motion of the reflecting elements would distort any wide band reflected signal unmercifully. Furthermore, its presence might prove an embarrassment to future generations wishing to make other use of the same space. It does promise a simple and immediate means of narrow-band service however.

(6) An isotropic active repeater (Fig. 2 (a)). An active repeater can interpose a large gain in the path, so that much smaller power need be radiated from the ground, and much less ground receiver sensitivity is required than for any of the passive repeater systems. The price for this, however, is the inclusion of active elements of unknown life and reliability, restriction in repeated signal characteristics, and a greater susceptibility to interference by other signals.

In the case of the active satellites the path is broken into two sections with relatively small attenuation in either part. The attenuation for each is:

$$\frac{P_T}{P_R} = \frac{4\pi p^2}{A\eta G} \tag{11}$$

For an ideal isotropic repeater G is unity, and in practice it may be as large† as 1/2. If polarization is not accounted for in a rather sophisticated way, another factor of 1/2, in average attenuation, and additional deep fades may occur. For this reason it does not appear desirable to use a completely unoriented active repeater.

(7) Oriented active repeaters. Antenna directivity in the satellite can be exchanged for r.f.

* Assuming all dimensions are very large compared to the wavelength.

† In the direction of weakest radiation.

power and sensitivity, or antenna gain for repeater gain. A highly directive satellite antenna obviously calls for very accurate sensing and control of orientation, and an astute study of the system is required in order to balance the requirements economically. There are three unique possibilities, however, which serve as guide posts in early system considerations such as presently concern us. They are:

(a) Spin stabilized repeater. A spinning satellite has a fixed axis which may be oriented perpendicular to the plane of the orbit as shown in Fig. 2 (b). If then a dipole radiator is directed parallel to this axis, there is a power gain of 3/2 to be had. More directive annular radiators such as a multiwavelength broadside array or a biconical horn also may be used, but since the ground stations are not generally directly below the satellite one would not necessarily benefit from high gain.

(b) Complete Earth illumination, or "broadcast" repeater (Fig. 2 (c)). The satellite may advantageously have just sufficient directivity to illuminate the visible part of the Earth. It would not necessitate being directed to a particular ground station, and would be useful for broadcast type coverage. An empirical relation between gain and directivity of a practical antenna is found to be

$$\text{Gain} \cong \frac{7{\cdot}5}{\theta^2} \tag{12}$$

At the worst position, a ground station will be at the edge of the beam and realize only half this gain. The angle subtended by the Earth at a 24-hr satellite is

$$\sim \frac{8000}{26{,}300} \cong \frac{1}{3} \text{ rad} \tag{13}$$

Thus, the maximum gain realized at the Earth horizon within this limitation is 33·7, or 15 dB from an antenna having 18 dB axial gain. Higher gain antennas would require directing to particular geographical areas.

(c) "Balanced" design (Fig. 2 (d)). This is a repeater in which the weight and complexity of the antennas are commensurate with the weight and complexity required to give the required output power. The balance calls for high gain antennas which in some cases must be independently directed toward the terminal stations.

The economic balance between power and weight in the satellite is almost entirely a competition between antenna size and weight and the

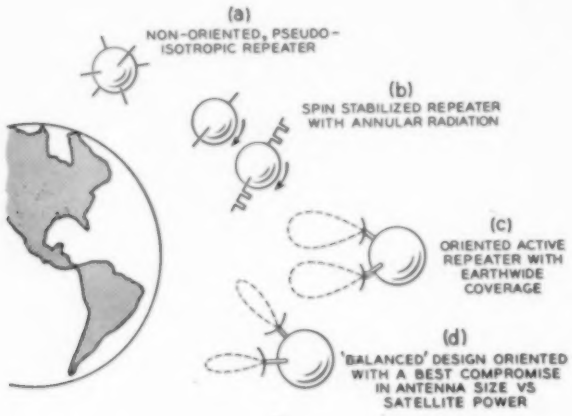

(a)
NON-ORIENTED, PSEUDO-ISOTROPIC REPEATER
(b)
SPIN STABILIZED REPEATER WITH ANNULAR RADIATION
(c)
ORIENTED ACTIVE REPEATER WITH EARTHWIDE COVERAGE
(d)
'BALANCED' DESIGN ORIENTED WITH A BEST COMPROMISE IN ANTENNA SIZE VS SATELLITE POWER

Fig. 2. Several forms which an active satellite relay may take.

weight and reliability of radio frequency power sources.[4] Determining this balance is a serious problem in system design, but it is worthwhile to look at the order of magnitudes indicated by present knowledge. We therefore calculate the following example based upon best guesses as to the weight of the largest repeater components.

At present, it appears that solar power can be harnessed at a cost of 1·6 lb/W of power by using oriented solar cells plus a nickel–cadmium storage battery.[5] If this is converted to radio frequency energy at a 10 per cent efficiency, we require about 16 lb of power supply per watt of r.f. power. The use of non-oriented solar cells would require at least 4 times as much weight.

It is necessary to make some estimate of the weight of a practical satellite antenna. It appears possible to make a practical antenna for space requiring only 0·1 lb/ft² of reflector area,* or a weight 0·14 λ^2 times the antenna gain (λ being in meters, and area efficiency being 60 per cent).

A satellite repeater for two-way communication using high directivity will presumably have two transmitters and two antennas, the receivers being duplexed on to the same antennas. On this basis there is an optimum antenna area of 160 ft²/W of radiated power. If non-oriented solar cells are used, 640 ft² of antenna/W of r.f. power would be justified. At low power levels and ultra-high frequencies, these are rather ridiculous proportions and must be modified to account for orientation problems and antenna tolerances. However, it is clear that if the complication of orientation control for point to point communication is accepted, one will have very large antennas before high satellite power can be justified. Since G and P_T in equation (11) are in constant proportion, the required satellite power and weight vary directly proportional to the satellite range, instead of proportional to the square of range as is the case for satellites having a fixed antenna gain.

* This would hold for a paraboloid made of 0·010 in. magnesium sheet, or 0·020 in. × ½ in. steel wire mesh, or a 0·005 in. thick metalized mylar balloon-paraboloid. Conceivably these could be made even lighter than here assumed, but there is also a problem of packaging and erecting the antenna. It is expected that future development will make both the antenna and battery weights quoted seem extravagant.

ATTITUDE CONTROL

Any satellite repeater which makes use of directional antenna characteristics requires some system of stabilization. The alternatives are the spherical (balloon) reflector, which is completely non-directional, and the pseudo-isotropic radiator. The latter does not provide a very acceptable solution, because either the radiation is far from isotropic, or else the polarization of the radiation is different in both sense and ellipticity in different directions, and can be used efficiently only with very complex methods of polarization tracking or diversity.

The spinning satellites of Figs. 1 (e) and 2 (b) are controlled by the gyroscopic action due to their own angular momentum. The guidance and control required to start a satellite spinning accurately on an axis perpendicular to the plane of its orbit does not look difficult. To assure that the axis will stay so aligned, and to maintain the spin is another question. It has been observed that the early Sputniks and Explorer satellites lost angular momentum at an observable rate, and also turned to spin about a different axis, presumably due to eddy currents produced by the Earth's magnetic field, atmospheric friction and/or the motion of loose wires, etc. in the gravitational field. Also, an inclined orbit plane precesses due to the Earth's bulge, so that the relative orientation of the axis of spin with respect to the orbit plane changes at a significant rate unless the orbit is very precisely over the poles or is precisely equatorial. It is clear that decelerating forces must be minimized, and it may be necessary to provide for a source of energy to maintain the angular momentum and to adjust the orientation of the axis of rotation of spinning satellites. Small vanes projecting from the satellite, black on one side and reflecting on the other, can be used to gather angular momentum from the Sun's radiation and/or from the momentum imparted by the heat radiation from the satellite. These two "radiometer" effects are in opposite directions and nearly balance unless the radiation balance

is destroyed. This may be accomplished by using surfaces with non-uniform spectral emissivity, or by heating or cooling the vanes to other than their equilibrium temperature. The latter may usefully be a factor in controlling the heat balance of the satellite. Fortunately, most foreseeable applications of such spinning satellites are at relatively low altitudes (not more than a few thousand miles), where the Earth's magnetic field is significantly large and has polar symmetry and therefore can be used to stabilize the rotation and to orient the satellite axis, making the attitude control problem of spinning satellites relatively simple. A possible expedient, for instance, is to equip the satellite with a small audio oscillator which drives a series of coils to create a rotating magnetic field. This interacts with the Earth's field to spin and orient the satellite much like the armature of synchronous electric motors.

In the case of the directively radiating satellites of Figs. 1 (f) and 2 (c) it is necessary to keep the whole satellite oriented so that the returned beam is pointed toward the center of the Earth. A number of means, such as servo systems controlled by horizon detection, are possible. Alternatively the attitude can be measured accurately by radio means and control instructions given, using very narrow-band telemetry channels, the bulk of the control apparatus being on the ground. The mechanical control may be provided by reaction from the angular momentum of a pair of flea-power motors requiring only a fraction of a watt of power for control of very large satellites. The system requires no gears or pulleys or reference masses or gyroscopes since the reaction to the armature motion alone is sufficient to rotate the body of the satellite in the vacuum of space.

In order that the satellite does not accumulate an embarrassing amount of angular momentum from meteoric bombardment or whatnot, which would require continuous correction by the attitude control motors, a few large diameter wire hoops may be unfurled from the satellite and, by coupling to the Earth's magnetic field, produce a damping force to inhibit the accumulation of angular momentum.

It is anticipated that the orientation and control gear can be made small enough so that several duplicate equipments can be provided in each satellite in order to provide high reliability.

The very directive antenna arrays of Fig. 2 (d) require separate and very accurate orientation. In addition the complexity of this repeater probably makes it justifiable to orient the solar battery or other thermal engine toward the Sun. Altogether then, three separate orientations must be controlled. This may also be done by radio from the ground, or alternatively by sensing means in the satellite rather than on the ground. This subject certainly requires a great deal of study before a long life reliable complex 24-hr satellite will be feasible.

MODULATION SYSTEMS

To make proper use of the radio channel which the satellite offers requires that some thought be given to the modulation system employed. A simple amplitude modulation system is extravagant of power, whereas a pulsed or an FM system uses more bandwidth. Wide band systems are more compatible with frequency sharing, however, and actually promise more economical spectrum utilization than narrow-band systems.

It is not clear what is the optimum modulation system for satellite communication. Work being carried out at present at Bell Laboratories will help to clarify this point, and further work for any specific system is clearly called for. As a basis for present system calculations it appears that FM with a system of "correlation detection"* in the receiver, offers nearly as good a power efficiency, at moderate output signal-to-noise ratios as has been demonstrated.

According to information theory it requires a received power of at least 0·695 kT W/bit sec to transmit information. The above-mentioned system theoretically comes within a factor of 2 of accomplishing this and actually has been operated within a factor of 4 of the above limit.

* Alternatively called "FM with feedback", "phase lock detection" or "frequency tracking". The use of this system was suggested by W. M. Goodall and C. L. Ruthroff.

Such a system requires a received carrier-to-noise ratio, normalized to a base band bandwidth, of $C/kTB = 32$ to transmit a signal with an FM index of 20 giving an output signal-to-noise ratio of 40 dB. The r.f. bandwidth required is 40 times the base band bandwidth, but the r.f. power required is only one three hundredth that of a single side band system of equal quality, and interference is greatly reduced.

An active satellite repeater may simply amplify and retransmit the incoming r.f. signal on a new frequency, or completely demodulate and re-modulate the r.f. carrier. In either case some noise will be added which will degrade the quality of the received signal. If the transponder operates with a r.f. signal-to-noise ratio of 20 dB, it will add negligibly to the noise in the output. If a demodulating repeater is used, the same quality requires a signal-to-noise ratio of 20 dB in the i.f. band, $\frac{1}{20}$th as wide, an advantage of 13 dB. The signal-to-noise ratio requirement in the satellite receiver can be reduced at the cost of a greater required signal-to-noise ratio in the ground receiver, but output power in the satellite is so much more of a limitation than is the satellite sensitivity, that this is not an acceptable exchange. Alternatively, the added noise may be compensated for by increasing the FM index by about 40 per cent.

Many of the signals to be transmitted involve a good deal of redundancy, and here is a place where bandwidth narrowing schemes, such as differential systems of modulation, or the vocoder may be used to great advantage.

NOISE

New advances in the microwave art have reduced the problem of radio noise at microwave frequencies by orders of magnitude. Whereas microwave systems were notably noisier than VHF systems, with the application of the maser and parametric amplifier, the reverse is the case. At the present time, masers exist which give effective noise temperatures well below 10°K, at frequencies in the microwave range, and this is a hundredfold better

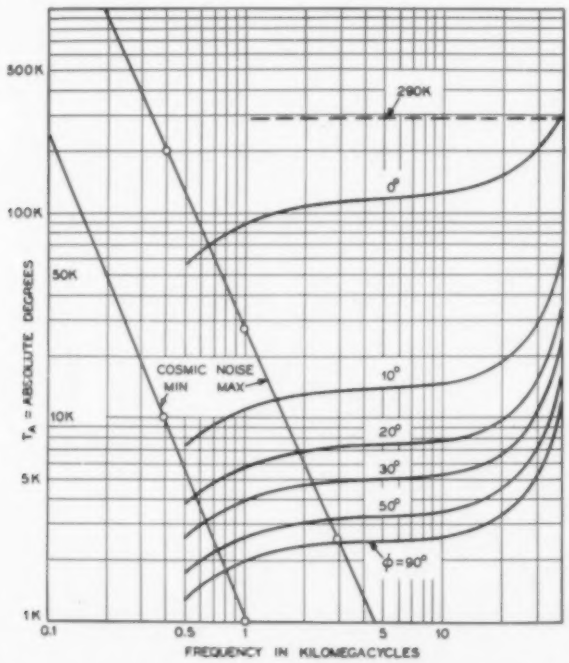


Fig. 3. Temperature of the sky as limited by cosmic radiation and atmospheric thermal radiation (oxygen absorption). From Ref. (2).

than is conventionally in use. Parametric amplifiers exist which give noise temperatures of as low as 100° in the same range.

These temperatures are less than some of the sky temperatures measured in radio astronomy.[6, 7] However, ordinary microwave antennas allow some spillover of energy to the side (or sensitivity to signals from the side) which contributes to the temperature seen by the antenna. This can account for at most a few tens of degrees, and with the use of a horn reflector antenna can be reduced to a few degrees.[8] The combination of the maser in a carefully designed antenna, with a very short waveguide connection makes it possible to realize very nearly a system noise temperature equal to that of the sky in the direction of the satellite.

Figure 3 shows the sky temperature as a function of frequency. In order to realize these very low temperatures one must avoid looking in the direction of the Sun (10^5 degrees at 2000 Mc/s) and the Moon (200°K). It is reasonable to design the system to take advantage of the very low sky temperatures, because coincidence of satellite direction and the Sun or Moon or even the galaxy center[9] direction will be uncommon, predictable and of very short duration.

Large antennas having directivities of a fraction of a degree are more subject to these hot areas than are small less directive ones which necessarily enclose a large area of cold sky in the beam. For this reason the ultra high frequencies look much more promising for large emplacements, and the very high frequencies (down to 40 Mc/s) look less unfavorable for small less directive (mobile) applications.

In the following calculations it is assumed that the antenna is kept pointed above $7\frac{1}{2}°$, to avoid the increased temperature due to oxygen absorption and ground proximity. Thus a receiver plus sky temperature of 20° is assumed at frequencies above 1000 Mc/s.

The satellite receiver, looking at the Earth, will see about 300° anyway, but for reasons of reliability, simplicity and to give relative freedom from interference it seems best not to plan on a noise figure of less than 10 dB.

MUTUAL VISIBILITY AND THE NUMBER OF SATELLITES REQUIRED FOR CIRCUIT RELIABILITY

In order to have a reliable communication path, there must be a high probability of there being a satellite mutually visible from each terminal, at an angle sufficiently above the horizon to avoid serious path degradation due to atmospheric refraction, absorption and thermal radiation. Thus it is necessary to define the region in which the satellite is visible from both terminals and to determine how many satellites are necessary to maintain a reasonable circuit reliability.

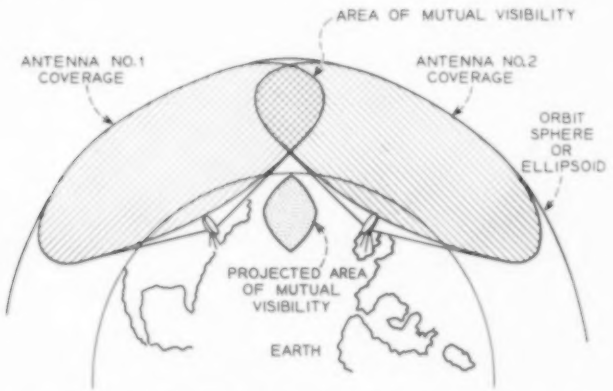


Fig. 4. Area of mutual visibility and its ground projection.

The acceptable positions of satellites for a good quality communication path are bounded

roughly by the volume common to two right circular cones centered vertically over the two terminal stations, as shown in Fig. 4. The intersection of the cones on the ellipsoid surface described by the satellite motion with respect to the surface of the Earth defines the area of mutual visibility.

It is convenient, in considering the mutual visibility problem, to refer to the vertical projection of this area on the surface of the Earth.

From these relations we can compute the maximum acceptable slant range and ground range for various satellite altitudes.

The relative goodness of the satellite relay path depends critically upon the elevation of the satellite (ψ) as seen at the receiver, because of the variation of the sky noise temperature with elevation, as can be seen in Fig. 7. Because of this and the problem of interference with terrestrial services it clearly is not economic to require the system to operate all the way down to the horizon. Since we do not yet have a better evaluation of these limitations, we arbitrarily limit our consideration to elevations at which the sky noise is not seriously limiting. An elevation of $\psi = 7\frac{1}{2}°$ above the horizon is a convenient value and will arbitrarily be taken as bounding the useful area of mutual visibility. A more astute study of path quality will undoubtedly change the shape and size of this area.

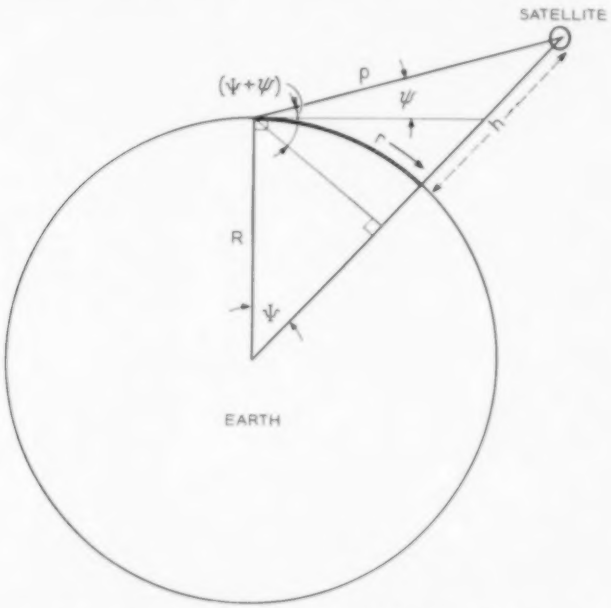


Fig. 5. Geometry for the computation of slant range and ground range.

In other words we will refer to that area of ground which the satellite is above when it is visible at both terminals.

If the satellite is in a circular orbit at altitude h and the acceptable elevation is ψ or more above the horizon, then from the simple geometry in Fig. 5 we get the following parametric equations giving p and r, the maximum range of the satellite and the projection of that range on the ground:

$$h - R(\sin \Psi \tan (\Psi + \psi) + \cos \Psi - 1) \quad (14)$$

and

$$p = R \frac{\sin \Psi}{\cos (\Psi + \psi)} \quad (15)$$

$$r = R\Psi \quad (16)$$

The area of Earth common to two circles of radius r drawn from the two terminal stations of a satellite relay path describes the part of Earth over which the satellite is acceptably visible from both terminals. This is the projected

area of mutual visibility and is roughly the area, if not the exact shape, of the region over which the satellite can be and provide acceptable relaying of messages between the two terminals. No matter how many satellites are erected, if their phase in orbit is random, there will be times when all satellites are outside the area of mutual visibility. These times may be made rare and of short duration, by erecting enough satellites, and in any case are predictable. Some day, perhaps, it will be possible to erect the satellite with sufficient precision to control the orbiting velocity and avoid this problem. Meanwhile it is well to consider how many satellites are required if the orbit phase is random.

The probability of there being a satellite in the area of mutual visibility is a function of the number of satellites, the orbit geometry and the distribution of satellites in orbit.

Let us assume that there are n satellites, distributed in m orbits, each orbit being in a plane having the same inclination with the Earth's axis and making the same angle with its neighboring orbital planes. There are thus n/m satellites in each orbit. The many orbits outline part of a sphere concentric with the Earth. The orbits are regularly spaced by accurate control of the launch time, and are assumed to be close together compared to the width of the area of mutual visibility.

Now consider an elemental area of this orbit sphere which is described by a parallelogram having a width $R\pi/m$ equal to the separation of orbit planes in longitude, and $R\mathrm{d}\theta_1$, a small fraction of the orbital circle. The probability that a particular satellite in the intersecting orbit lies in the area so defined is simply:

$$\frac{\mathrm{d}\theta_v}{2\pi} \tag{17}$$

and the probability that it lies outside is of course:

$$1 - \frac{\mathrm{d}\theta_v}{2\pi} \tag{18}$$

The probability that all the n/m satellites in the orbit are outside the area is

$$\left(1 - \frac{\mathrm{d}\theta_v}{2\pi}\right)^{\frac{n}{m}} \tag{19}$$

Now if Φ_2 is the angle of longitude there are $N = (m/\pi \mathrm{d}\theta_v) \iint_{\mathrm{MV}} \mathrm{d}\theta_v \mathrm{d}\Phi_v$ such elements in the area of mutual visibility (MV). Assuming $\iint_{\mathrm{MV}} (\mathrm{d}\theta_v \mathrm{d}\Phi_v / 2\pi^2) \ll 1$ (i.e. that the absence of a satellite from any part of the area does not significantly increase the probability that it is in the remainder) there is a probability,

$$P \simeq \left(1 - \frac{\mathrm{d}\theta_v}{2\pi}\right)^{N\frac{n}{m}}$$

$$\simeq \left(1 - \iint_{\mathrm{MV}} \frac{\mathrm{d}\theta_v \mathrm{d}\Phi_v}{2\pi^2}\right)^n \tag{20}$$

that there is no satellite in the area at any particular moment. The above approximation holds only if the area of mutual visibility is a small fraction of the celestial sphere, which is the case for any practical earthbound application.

The double integral is equal to the area of mutual visibility multiplied by a function of the latitude of the area and the inclination of the orbit, and is not critically dependent upon the shape of the area except in some peculiar extreme cases. Thus equation (20) may be written

$$P \simeq [1 - A_{\mathrm{MV}} F(L, Q)]^n \tag{21}$$

where A is the projected area of mutual visibility and $F(L, Q)$ a function of the latitude of the path and the orbit inclination.

The foregoing process has been carried out for some particular cases of interest and the results are shown in Figs. 6 and 7. It is clear that inclined orbits have a moderate advantage only for low orbits at moderate latitudes of only 20° or 30°. Otherwise, polar orbits are best except for terminals close to the equator, where equatorial orbits are most advantageous.

The number of satellites required is a severe function of satellite altitude. However, the cost of erecting satellites increases rapidly with altitude, both because the satellite must be larger to account for the additional path attenuation and because of the additional propulsion requirement. There is therefore an optimum altitude for communication satellites, as we shall soon see.

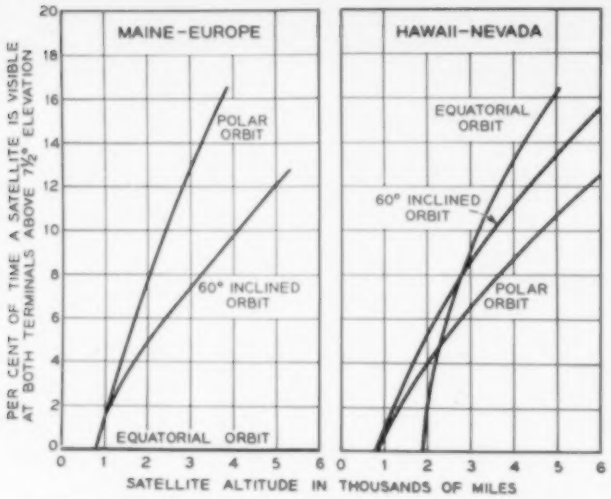


Fig. 6. Mutual visibility times for various satellite orbits, for a 3000 mile ground range.

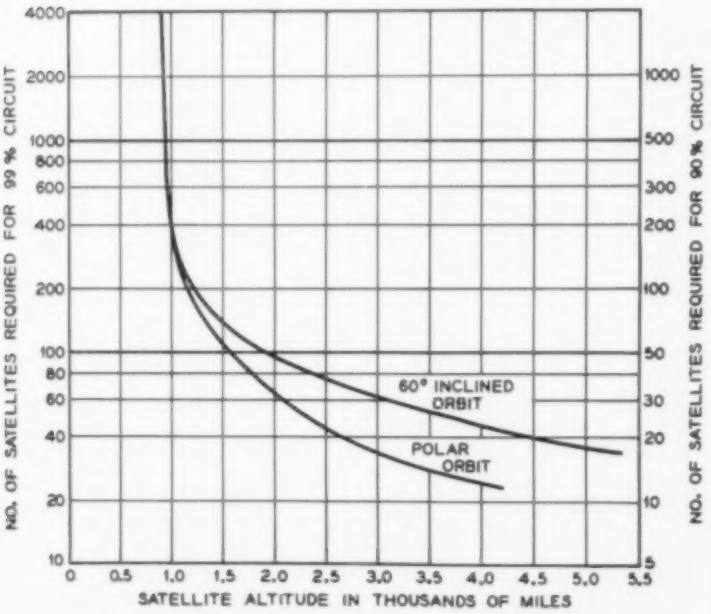

60° INCLINED ORBIT
POLAR ORBIT
NO. OF SATELLITES REQUIRED FOR 99 % CIRCUIT
NO. OF SATELLITES REQUIRED FOR 90 % CIRCUIT
SATELLITE ALTITUDE IN THOUSANDS OF MILES

Fig. 7. Number of satellites needed for substantially continuous service between ground terminals 3000 miles apart. In this case the path was from northern U.S.A. to central Europe.

OPTIMUM ALTITUDES FOR RELAY SATELLITES

The best altitude for a satellite relay system is that which provides the most service for the least cost. All other satellite costs are dwarfed by the prodigious expense of putting the satellite into orbit. This can be expected to have a component due to the cost of hardware such as control and guidance equipment which are insensitive to satellite size and altitude, and a

component roughly proportional to the amount of propellant used. The cost function then must be something like

$$\text{Annual cost} = \frac{M}{L}(\alpha + \beta\, WF(h)) \qquad (22)$$

where M is the number of satellites per ground link, L the life, W the weight and $F(h)$ a function of the altitude of the satellite due to the cost of propellents, etc. The size of the constants α and β is undetermined. At present they must be especially large in order to account for development costs. Ultimately, when development has been written off, the cost of erecting satellites should be more reasonable than now. The function $F(h)$ is intended to account for the fact that the cost of putting a satellite in a circular orbit is a function of the height of the orbit. The economics of the system depend critically upon this characteristic, which we can do little better than guess at. Presumably, when satellites are erected in quantity, the factor α will be negligible and the cost of erecting satellites will be nearly proportional to the total weight of propellant used.

An idealized relationship between the velocity added by a rocket stage[10] is

$$v = c \log_e \frac{m_0}{m_1} \qquad (23)$$

where m_0/m_1 is the ratio of the total weight of the stage, including its payload, to its burned out weight, and c is the effective jet velocity.

A fraction of the weight of each stage consists of burned out rocket motor and tanks, which are no longer useful. In the past this has amounted to about 20 per cent of the initial weight of the stage, and there is now promise that it may be made as low as 5 per cent. Thus if the burned out rocket and guidance equipment without payload accounts for a fraction of the initial weight of a single stage:

$$\frac{\text{Ratio of burned out stage weight*}}{\text{to initial weight}} = (1 - \eta) \qquad (24)$$

* Including weight of guidance and control equipment which are not part of the useful payload.

the useful payload weight ratio per stage

$$\frac{M_{(K+1)}}{M_K} = \frac{\text{Ratio of useful payload weight}}{\text{to initial rocket weight, per stage}} = \left[\exp\left(-\frac{v}{c}\right) - (1 - \eta)\right]. \qquad (25)$$

The over-all weight ratio M_K/M_0 for a rocket of several stages is found by multiplying several such factors, one for each stage.

This has been done for several cases, for various values of c and η, and a trajectory approximating the ideal program for erection of a satellite, using two stages to get the satellite into a transfer ellipse, and a third vernier stage to convert to a circular (polar) orbit. An additional 0·5 km/sec was arbitrarily included in the calculation to account for losses due to atmospheric friction and gravitational acceleration during the finite firing time. The upper part of the shaded area of Fig. 8 is believed to be in the range of demonstrated performance, and the lower part covers the range of optimistic prognostication for the next few years. It is clear that rough conclusions drawn on the shape rather than the magnitude of the mass ratio curve will be valid regardless of foreseeable developments in propulsion efficiency.

Now we have a rough evaluation of the several factors which vary with satellite altitude. The total mass of satellites required is dependent upon the range and the mutual visibility. Combining these "known" factors we have

$$M_0 = \text{Weight of rocket required} = M_1 n \left(\frac{M_0}{M_1}\right) \qquad (26)$$

where n_1 and M_0/M_1 are given as a function of altitude in Figs. 7 and 8. According to the path loss equations, M_1 is proportional to d^4 for doubly curved reflectors, or d^2 for planar reflectors and some active satellites. In the case of the "balanced" active satellite design, the plane reflector, or the saucer reflector, the required weight increases linearly with range. Since we already have range vs. altitude from Fig. 4, we are in a position to plot relative rocket weight vs. altitude of satellites. This is done for the case of a 3000 mile path over the North

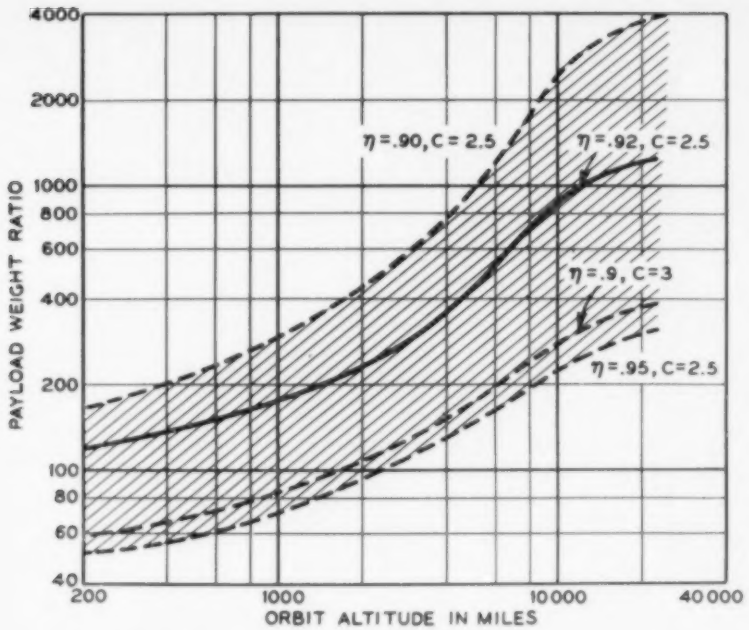


Fig. 8. Payload weight ratio for various rocket characteristics.

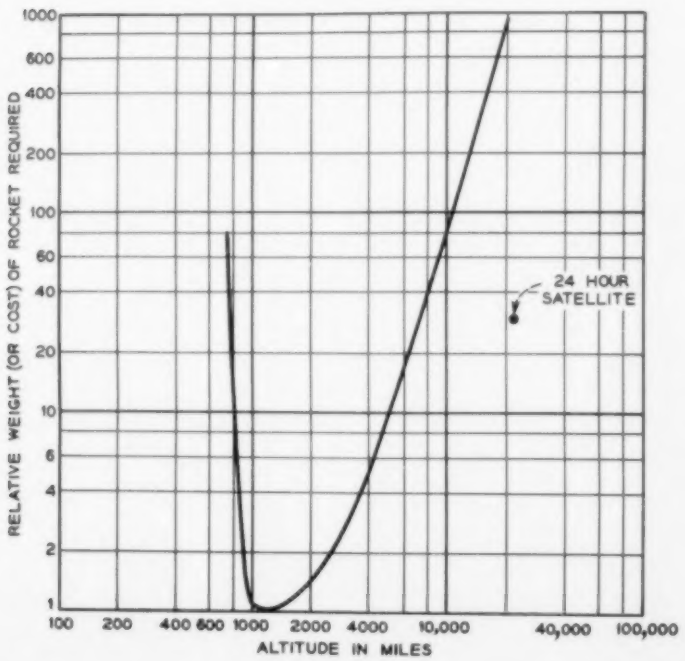


Fig. 9. Relative rocket propulsion weight required to establish spherical passive satellites for communication over a 3000 mile path.

Atlantic. It is evident from Figs. 9 and 10 that there is an optimum altitude of about 1200 miles for passive repeaters and 2000 miles for active satellite repeaters, and 22,300 miles for active satellites of the "balanced" design, or for a saucer or plane reflector giving a particular ground area coverage. A shorter path, such as the 2000 mile Newfoundland–Hebrides path has a much lower optimum altitude.

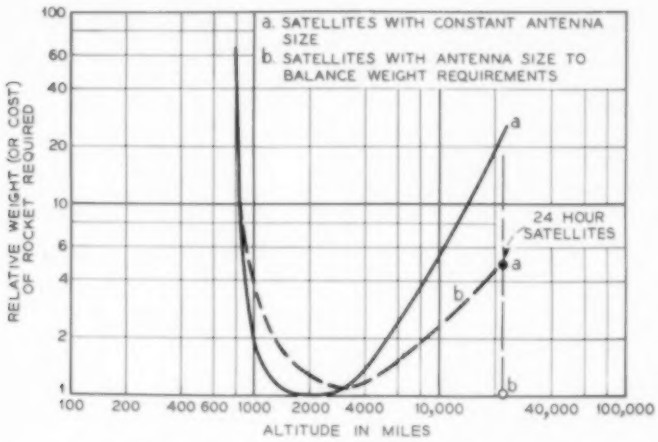


Fig. 10. Relative rocket propulsion weight required to establish an active repeating satellite system for communication over a 3000 mile path.

More or less efficient rockets will have little effect upon the shapes of the curves, but a very large effect upon the scale factors. These considerations lead one to believe that in the case of the simple repeaters it is most efficient to use many, relatively low satellites rather than the very high satellites sometimes proposed. However, using many low satellites increases the tracking and guidance problem, since visibility times are reduced and satellites pass by faster and more often. The more complicated satellite configurations can only be justified in the higher, 24-hr orbit, which is unique in that only one relay is required for a transoceanic path, and satellite tracking is greatly simplified or even eliminated.

GROUND STATION

What kind of ground station should one use for satellite communication? Prophets on this subject have predicted everything between a watt and non-directive antennas to hundreds of kilowatts with colossal antennas. Even though we are not in a position to make a proper system evaluation, we should be able to arrive at the order of magnitude of the economic balance between satellite and ground station proportions.

The economics relating fixed ground installation size and satellites must be commensurate with the economics relating rocket propulsion and payload. While one ground station may survive many satellites it is equally true that a satellite may serve many ground stations. At any rate, with large propulsion to satellite weight ratios, and being faced with the problem of operational reliability in the satellite, we can surely justify a ground station cost commensurate with putting satellites into orbit.

These considerations indicate that a few million dollars may be justified as the cost of a ground installation commensurate with satellites weighing a few tens of pounds in orbits at an altitude of a thousand miles or so. It seems practical, therefore, to use an antenna area in the order of 10^4 ft^2, and a total power consumption of hundreds of kilowatts.

In the case of a non-oriented passive satellite

(balloon) one is forced to use very large antennas and high powers in order to use a satellite of realizable size.

The oriented plane or oriented doubly curved satellite reflectors are much smaller for a given path loss, and justifies using a smaller ground installation.

In the case of active repeaters the frightful cost of power in the satellite still makes it desirable to use large ground receiver antennas and low noise maser pre-amplifiers. Amplification in the satellite is relatively cheap, however, and low power, small ground transmitter installations may reasonably be used, provided a sufficient signal strength is available to override any probable interference at the satellite receiver.

Mobile locations necessarily are limited to small structures and low powers, by operational rather than economic consideration. As a consequence they require relatively large power and gain in the satellite repeaters. They are more subject to interference and thus can make much less efficient use of the medium by sharing frequency space with other services.

Since information providing a reasonable basis for proportioning the large elements of the system is so meager, we will simply determine what are consistent parameters for a number of possible situations.

DESIGN MARGINS

It is always a question as to how much margin should be allowed in system considerations of this kind. A 30 to 40 dB fading margin is necessary in most radio systems to account for deep fades and other interferences, and for equipment degradation. It is not certain how much margin should be allowed in communication via satellites, but it is likely that the usual causes of signal degradation will be conspicuously absent. Fading due to tropospheric variations should be very small because the wave will not be traveling parallel to the natural atmospheric stratification. Tropospheric scattering removes only a minuscule part of the energy from the beam giving a negligible attenuation of the transmitted beam. Atmospheric refraction is small, predictable, and relatively constant. Ground reflection interference should be absent because the highly directive beams directed upwards stand no chance of illuminating the ground significantly. Ionospheric effects are reduced to an insignificant rotation of polarization of the ultra high frequency waves. Radio star scintillations are observed at much lower frequencies, presumably due to ionospheric effects and so portend no trouble. There is an absorption due to oxygen and water vapor in the atmosphere but this is negligible below about 20,000 Mc/s. In short, the bothersome terrestrial effects all are small, and the wave travels almost entirely in the same ideal medium that propagates light and radio waves from the stars. It appears that radio is finally coming into its own, the inverse square law can be applied with impunity, and that no margin need be allowed for propagation effects.

Some allowance must of course be made for equipment degradation. In the calculations which follow, no such margins have been allowed except in the use of empirical instead of theoretical relations regarding antenna gain, directivity, modulation efficiency and power generation efficiency. This is done for two reasons, (1) so that the figures can be used as design objectives, and (2) so that margins may be knowingly applied to the results without losing their identity. System compromises for the sake of expediency may also call for additional margins.

It appears that an operating margin of only a few dB should be sufficient for the systems described. Of course this remains to be proved, and this is the important reason for conducting systems tests as soon as possible.

CONCLUSIONS

By making some very broad, order of magnitude assumptions as to the cost of the larger components of satellite communication system, and using the best indicated rocket propulsion and satellite weight factors, a very crude evaluation of system proportions has been made. Out of a very large number, a relatively narrow range of possibilities emerges as the most practical applications within the reasonably near future.

From the point of view of rocket propulsion it appears that an altitude of around 1000 miles is an optimum for passive balloon reflectors, and 2000 miles for modest sized active repeaters and ~22,300 miles for active repeaters using directive antennas and for plane or saucer reflectors. Because many fewer repeaters may be used, it becomes practical to think of an active system of orientation control for satellites in the 24-hr (22,300 mile high) orbit. These factors make the propulsion weight requirement for a system using such satellites comparable to that required for a substantially continuous circuit provided by lower satellites. The advantages that accrue from the nearly stationary aspect of the 24-hr satellite may well compensate for the complication in guidance necessary to place and maintain them accurately in orbit.

The higher microwave frequencies, up to 10 kMc/s, are propitious for broadband applications using large ground installations and passive satellites because of the low sky noise temperature. There is a broad optimum around 2000 Mc/s for active repeaters having relatively small antennas. For repeaters with large antennas, optimized in size to give the smallest satellite weight, 10,000 Mc/s is best. For narrow band mobile services where it is less practical to use such low noise receivers it may be better to use lower (VHF) frequencies and reduce the satellite tracking problem. Some of the more practical cases, with a 10-dB design margin added, are:

(a) Point to point 5-Mc/s bandwidth service with a 40 dB output signal-to-noise ratio using passive satellites randomly spaced at 1000 miles or so altitude. A few hundred $\frac{1}{4}$ mil aluminized mylar balloons of 85 ft diameter weighing about 50 lb each are required to give a 99 per cent reliable channel, with a *10* dB *margin* allowed for possible path and equipment degradation at the extremes of range. The ground station requirements are: 50 kW average power at around 6000 Mc/s radiated from 110 ft diameter paraboloidal antennas, and maser operation so that the sky temperature at $7\frac{1}{2}°$ above the horizon limits the receiver operation.

The ground antenna beam width is only 0·0023 rad (0·13°) so the antenna must be pointed to an accuracy of a few hundredths of a degree.

Each of the requirements looks difficult, but possible. It may be that the economics of satellite erection will allow erecting much larger spheres, thus relaxing some of the ground station requirements. The beauty of this repeater is an indefinitely long life and an almost unlimited bandwidth.

(b) A likely looking alternative to the above is the use of about 60 spinning active repeaters in 2000 mile altitude polar orbits. Each of these would radiate $\frac{1}{2}$ W at 2000 Mc/s from dipole antennas to give the same 10 dB margin in received power, and would require only 200 W from the ground transmitter. The weight of such satellites is mostly in the solar battery and is estimated at about 40 lb.

(c) A 24-hr satellite competitive with the above could be a saucer reflector, which would have to be oriented to reflect the signal between the appropriate continents. For frequency of 6000 Mc/s and a ground coverage of 1000 miles diameter, it would be 200 ft in diameter and might weigh as little as 300 lb. For a single path only one repeater is required, and three would suffice for world-wide coverage. Again the chief advantage is in an unknown but probably very long life compared to broad band active repeaters. Again, the unlimited bandwidth and ability to participate in many different circuits make it attractive as an ultimate facility.

(d) A most promising active 24-hr satellite would radiate $\frac{3}{4}$ W for the same 5 Mc/s intelligence bandwidth from an antenna having 18 dB gain. Without orientation of solar cells toward the Sun it appears possible that such a repeater could be built weighing as little as 100 lb. Orienting the solar batteries makes a more complicated structure but might reduce the weight to as little as 35 lb. The ground transmitter power could be as little as 200 W. This would be good for only one service at a time, it should be noted, additional services or broader bandwidths would require additional satellite weight in proportion. The complexity of the system makes the design for long life and reliability difficult.

(e) The foregoing repeater sizes could provide a single speech (3 kc/s) channel to and from smaller stations equipped with 20 ft antennas, and 0·02 times as much ground transmitter power, or teletype (30 c/s intelligence bandwidth) with 6 ft antennas and 0·002 times as much ground transmitter power.

The lower ground transmitter powers quoted are impractical in many cases because such operation leaves the satellite seriously open to interference from other ground stations.

This paper skirts a number of questions which require further serious investigation, most notable of which are:

(1) How much signal fading, if any, can be expected over the relatively ideal path to a satellite repeater?

(2) Can the shape of a large mylar balloon be maintained in space over long periods?

(3) Can the axial direction and spin of a spinning satellite be maintained for a long time?

(4) What is the best way to determine and control the attitude of a satellite over long periods?

(5) What are the optimum mutual visibility factors for a particular system signal-to-noise ratio?

(6) What are the actual propulsion factors for most probable rocket configurations, and what satellite altitudes does this indicate?

(7) What is the best modulation efficiency that can be realized in practice, and what is the best modulation system with respect to efficiency and the avoidance of interference?

(8) What sky noise temperatures are actually encountered in the microwave region, and how low system temperatures can be realized?

(9) Just what levels of interference from existing and proposed radiations can be expected at reasonable terminal locations?

(10) How good a discrimination against back and side radiation can be obtained in a giant antenna?

(11) What are more accurate economic factors of rocket propulsion and equipment costs in relation to the proposed systems?

(12) What are the possibilities of component reliability for unattended operation over long periods?

These and many other problems must be studied in much greater detail before a practical satellite relay system can be engineered. The promise in larger capacity, more economic transoceanic communications is so great as to justify extensive effort in this direction.

Acknowledgments—The author is indebted to J. R. Pierce, R. Kompfner and W. C. Jakes and other associates for many helpful discussions.

REFERENCES

1. A. C. Clarke, *Extraterrestrial Relays* in *Wireless World* p. 153 (Oct. 1945).
2. J. R. Pierce, *Jet Propulsion* **25**, 153 (1955).
3. J. R. Pierce and R. Kompfner, *Proc. Inst. Radio Engrs., N.Y.* **47**, 372 (1959).
4. R. S. Davies and C. S. Weaver, *Proc. Inst. Radio Engrs., N.Y.* **47**, 1151 (1959).
5. H. K. Van Ziegler, p. 62 in *Scientific Uses of Earth Satellites* (edited by J. A. Van Allen), University of Michigan Press.
6. D. C. Hogg, *Effective Antenna Temperature Due to Oxygen and Water Vapor in the Atmosphere*, to be published.
7. P. D. Strum, *Proc. Inst. Radio Engrs., N.Y.* **46**, 43 (1958).
8. R. W. Friis, *A New Broadband Microwave Antenna System* in *Communication and Electronics* p. 97 (March 1958).
9. R. H. Brown and A. C. B. Lovell, *Exploration of Space by Radio* Chapt. 4, Wiley, New York (1958).
10. C. C. Adams, *Space Flight* p. 74, McGraw Hill, New York (1958).

A STUDY OF UHF SPACE COMMUNICATIONS THROUGH AN AURORA USING THE MOON AS A REFLECTOR*

R. P. INGALLS J. C. JAMES and M. L. STONE
Massachusetts Institute of Technology, Lincoln Laboratory

Abstract—A lunar reflection experiment has been performed with a transmitting site in the auroral zone and two receiving sites at mid-latitudes. With this arrangement a study has been made both of the reflection characteristics of the moon and of the characteristics of UHF signals from space propagated through auroral-type disturbances. It was found that auroral ionization in the propagation path increased the signal fading rate and produced rapid fluctuations in the Faraday rotation but did not absorb an appreciable amount of UHF energy.

INTRODUCTION

Problems associated with space communications and the detection of ballistic missiles have stimulated interest in the study of UHF propagation through the ionosphere of the auroral zone. Prior to the availability of a polar transiting satellite several groups made studies of radiation from radio stars to deduce effects produced on a wave penetrating the aurora, while others probed the auroral disturbance with radars to investigate its back-scattering properties. Since a complete model of the aurora is not yet available it is difficult to make use of the results of these experiments to predict with certainty the effect of the aurora on the spectral bandwidth, polarization, and absorption of UHF energy.

The present study therefore investigated auroral effects by using a signal which was transmitted to the Moon from a site in the auroral zone and was received after reflection at mid-latitude sites, which are essentially free of auroral effects. On the basis of previous work on the measurement of characteristics of lunar echoes, carried out at mid-latitudes, certain anomalous features of the signals recorded in the present experiment could be attributed to the presence of ionization associated with aurora. The presence of an auroral disturbance in the transmission path to the Moon was verified by visual or radar observations made from the transmitting site. In order to use the Moon for these propagation tests a rather complete understanding of the reflecting properties of the Moon and the transmission characteristics of the normal ionosphere was necessary. This information was provided by previous lunar experiments and included the effective Moon radar cross section, characteristics of the signal fading, Doppler shift variations and polarization variations.

Radar measurements of the Moon were first reported in 1946 and since that time various experimental studies of radio wave reflections from the Moon have been made.[1–14] Some of these experiments were primarily directed toward measuring astronomical properties of the Moon by radio means while others simply used the Moon as a convenient reflector to aid in a study of the Earth's ionosphere. From these observations the following four characteristics of lunar echoes at UHF have been fairly well established.† (1) The effective lunar radar cross section for continuous waves at UHF is of the order of 5×10^{11} square meters, while the transmission loss has about a 2 dB variation due to the change in observer–Moon distance. Except for this small variation the average power

* The work reported in this paper was performed by Lincoln Laboratory, operated by Massachusetts Institute of Technology; this work was supported by the U.S. Air Force.

The work at Stanford Research Institute was sponsored by the Rome Air Development Center of the Air Research and Development Command under contract AF 30(602)-1762.

This work was performed at the Defence Research Telecommunications Establishment under Project No. PCC D48-38-01-11.

† See, for example, Ref. (12).

level back-scattered from the Moon remains essentially constant. (2) Short period fades with periods of the order of seconds are observed and are attributed to the effective libration of the Moon, i.e. to the fact that the surface of the Moon as seen by an observer on earth changes slightly with time. (3) There is a Doppler shift of the lunar reflected signal which is due to the Earth's rotation and to the Moon's elliptic orbit. The maximum Doppler frequency is of the order of hundreds of cycles per second for signal frequencies in the UHF band. The motion of the observer due to the Earth's rotation is responsible for the major component of this Doppler shift. (4) The Faraday effect causes a rotation of the plane of polarization of the received signal. This effect is a consequence of the propagation of a radio wave through an ionized medium in the presence of a magnetic field. The amount of rotation depends upon the electron distribution in the ionosphere and upon the magnitude of the component of the magnetic field in the direction of propagation. The amount of rotation changes continuously due to varying ionospheric conditions and due to changes in the orientation of the propagation path relative to the magnetic field.

The results of the experiment described in this paper are of two categories, namely those obtained when the signal traversed a quiet ionosphere, and those obtained when the signal traversed an auroral disturbance. Under quiet ionospheric conditions the results were essentially the same as those obtained at mid-latitude stations. Under aurorally disturbed conditions the 440 Mc/s lunar signal was perturbed in two respects: (1) the fading rate of the received signal increased, and (2) the polarization angle exhibited erratic fluctuations. However, there was no measurable attenuation of the signal during times of auroral disturbances. A description of these results will be given later along with estimates of the change of electron content of the ionosphere during auroral disturbances.

THE EXPERIMENT

The basic experimental technique was to transmit a CW signal at 440 Mc/s from a site in the auroral zone, reflect this signal from the Moon, and receive the reflected signal at two sites outside the auroral zone. The received signal was perturbed both by transmission through the Earth's atmosphere and by reflection from the Moon's surface. The determination of the degree to which these effects could be separated was one of the main aspects of the experiment.

Accurate predictions of the Moon's position and the Doppler frequency shift were made from astronomical data. In addition predictions of the "apparent Moon libration" were necessary since the spectral width of a normal lunar signal could be directly related to it. The "apparent Moon libration" is the libration measured with respect to a coordinate system moving with the observer. It may be considered as an apparent rotation of the Moon about an axis that is normal to the line-of-sight to the observer. Actually most of this effect is due to the motion of the observer on the rotating Earth rather than the motion of the Moon itself.

The transmitting site was located at College, Alaska, in the auroral zone. The antenna was an azimuth-elevation mounted, 60-ft parabolic reflector with a horizontally polarized feed. The transmitter was a 10 kW klystron, driven at 440 Mc/s by a signal with a stability of at least 1 part in 10^8 per day.

The receiving sites were at Westford, Massachusetts (Millstone site) and near Ottawa, Ontario (Shirley Bay site). Identical receivers were used at each station. The antenna at Westford was an 84-ft parabolic reflector on an azimuth-elevation mount, and was equipped with a cross-polarized linear feed. The antenna at Ottawa used a 28-ft parabolic reflector with a cross-polarized feed. This antenna had a polar mount. The beamwidths of the antenna at College, Millstone and Shirley Bay were approximately 2·5°, 2° and 5°, respectively. The two azimuth-elevation mounted antennas were positioned manually every 2 min, using predicted position logs computed from astronomical data. During this time interval the Moon moved in position only a fraction of the antenna beamwidth. The polar mounted antenna at Ottawa

used its sidereal drive with small periodic corrections.

Position logs were prepared by means of an IBM 709 Computer program, which converted values of declination and Greenwich hour angle obtained from the *American Nautical Almanac* to values of elevation and azimuth at the particular site.

Fig. 1. The tuning was accomplished by the manual adjustment of a local oscillator. The frequency of this local oscillator and the time at which it was measured were recorded automatically. The frequency stability of the receiving system was of the same order of magnitude as that of the transmitter. Thus on a short-time basis the absolute Doppler shift was measured

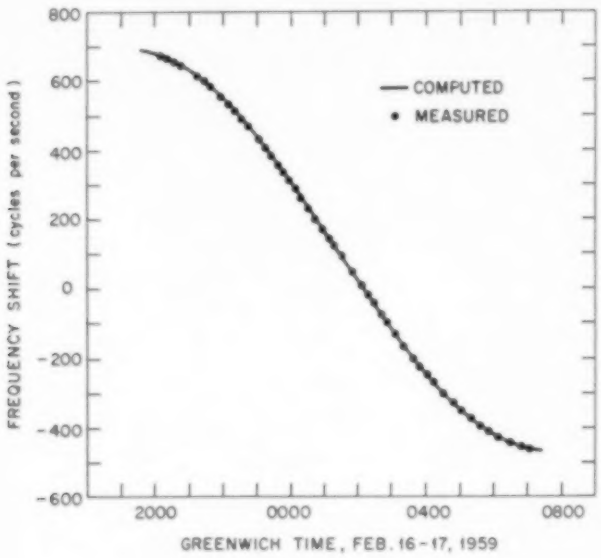


Fig. 1. A typical plot of the computed and measured Doppler frequency for Millstone.

Lunar transmissions were made for 20 min of each half hour. During the other 10 min a radar search at 398 Mc/s was carried out at the transmitting site while at the receiving sites the equipment calibration was checked.

Each receiver used for the experiment was a dual-channel, triple-conversion, superheterodyne system with an ultimate bandwidth of 25 c/s obtained from a crystal filter. Paper chart records were made of the detected signals and of the fading rate. Magnetic tape records of the detected signals and low frequency i.f. signals were also made. The Doppler frequency shift of the Moon-reflected signal changed typically about 1100 c/s during a run, so that the receiver had to be tuned continually during the experiment. A plot of this Doppler shift is shown in Fig. 1. The tuning was accomplished by the manual adjustment of a local oscillator. The frequency of this local oscillator and the time at which it was measured were recorded automatically. The frequency stability of the receiving system was of the same order of magnitude as that of the transmitter. Thus on a short-time basis the absolute Doppler shift was measured to an accuracy of approximately two cycles per second. A comparison of the measured and predicted Doppler shifts for a typical run is shown in Fig. 1.

RESULTS

Early in the experiment it was shown that the measured Doppler shift agreed well with the predicted values. It provided a useful check on the system operation and was continued in use throughout the experiment for this purpose.

Two samples of the detected signal with essentially no post-detection band limiting are shown in Fig. 2. The two traces labelled *H* and *V* represent the outputs of the horizontally and vertically oriented antenna feed elements. A rapid fading of the order of 1 fade/sec can be

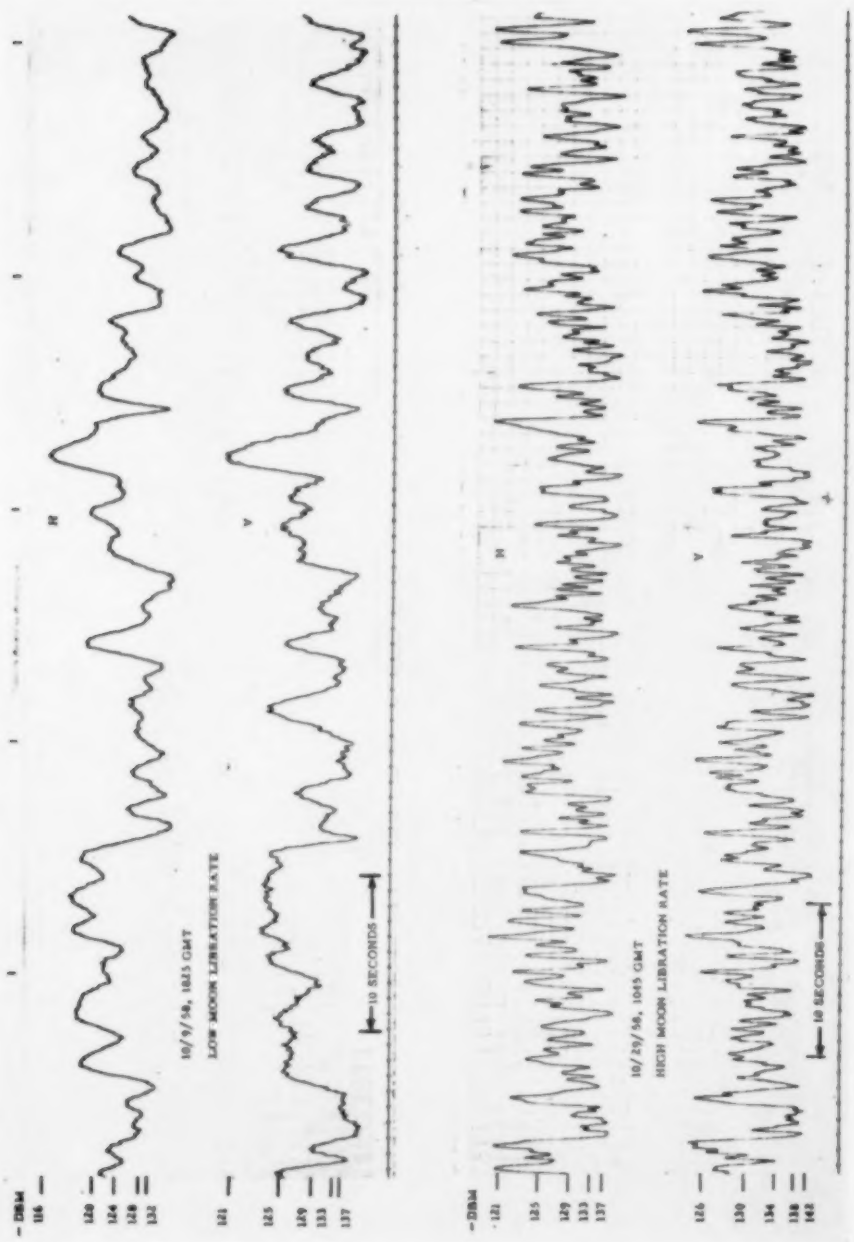


Fig. 2. Samples of the received lunar signal at Millstone as recorded on paper tape. The post-detection time constant was 0·1 sec. The *H* and *V* notations refer to the outputs of the horizontally and vertically polarized channels, respectively.

observed. This fading is interpreted as being primarily caused by the Moon's apparent libration.

Signal level measurements were made from detected signal records with a considerable amount of post-detection averaging. From the average signal levels in the two orthogonal channels the total power received at each of the two sites for 1 day's run is shown in Fig. 3, where each point represents the sum of 6-min

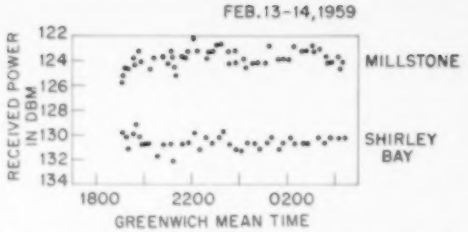


Fig. 3. A plot of the average total power received at each of the two sites throughout a typical lunar day.

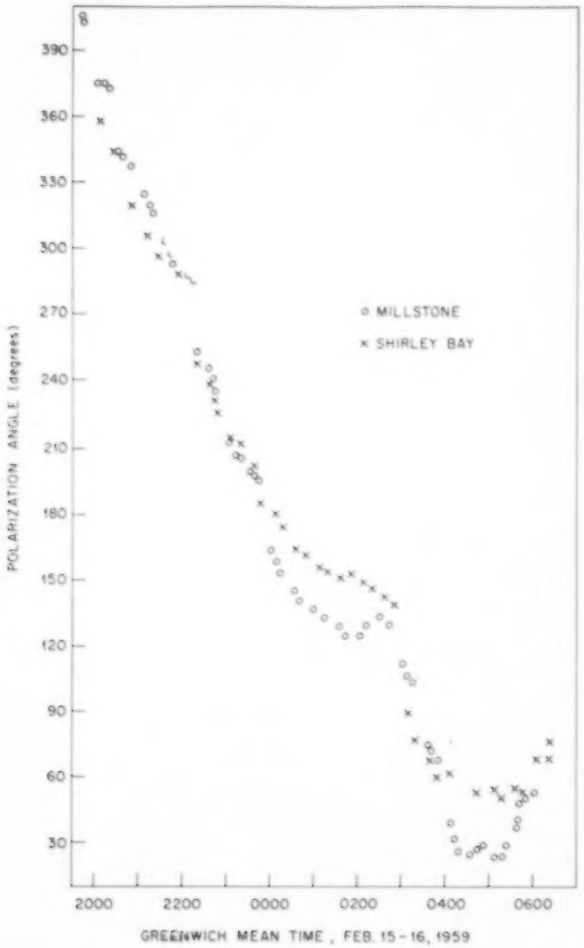


Fig. 4. The variations of the polarization angles at the two receiving sites for a typical lunar day.

averages of the power in the two channels. The difference in the levels of received power at the two sites is due to the difference in gain of the two receiving antenna systems. The average power of the lunar return for each day, when corrected for range, was constant to within

about 2 dB, which represents the order of accuracy of the power measurements. No significant reduction of the signal level was noted during any of the auroral disturbances which occurred during the tests.

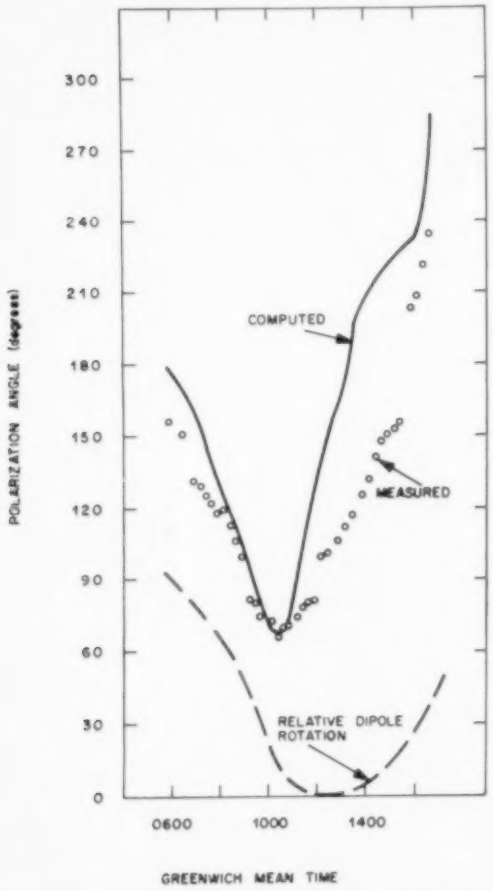


Fig. 5a. A comparison of measured and computed variations in polarization angle for the lunar signal received at Millstone on October 4, 1958. The variation in polarization angle due only to relative dipole rotation is shown by the dashed curve.

The radar cross section of the Moon computed from the average value of the received power was of the order of 5×10^{11} m², i.e. about 5 per cent of the projected visible cross section. This value of fractional cross section can be apportioned between a reflection coefficient and the gain of the Moon as a reflector. Knowledge of one must be obtained before the other can be deduced. There is strong evidence from short pulse measurements[7] and from spectral broadening of both CW[12] and long pulse signals[10] that most of the effective reflection takes place from a relatively small central disc on the Moon. More is to be said about spectral broadening in the present experiment.

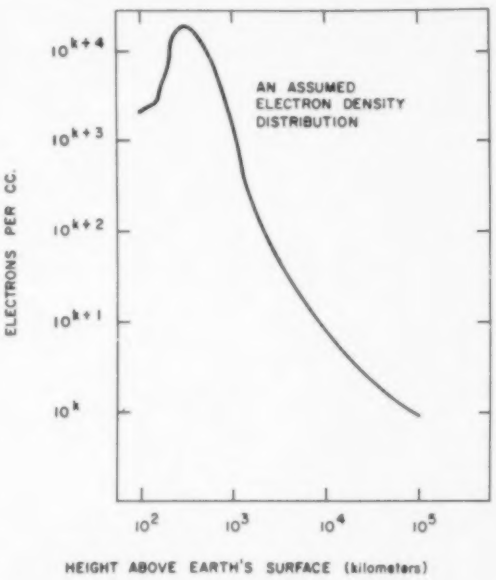


Fig. 5b. An assumed electron density distribution used in computing the predicted change in polarization angle. The absolute level is adjusted by varying parameter k.

The polarization angle of the received signal was defined as the arc tangent of the ratio of magnitudes of the average signal levels in the two channels. The quadrant ambiguity was resolved by observation of the reversal of phase of the i.f. signal as the plane of polarization rotated through an angle of 90°. The measured polarization angle is then ambiguous to a multiple of 180°. Approximate computed values enabled this ambiguity to be resolved in most instances.

The measured variation of the polarization angle of the received signal throughout a lunar day for both Millstone and Shirley Bay is shown in Fig. 4. The values for Shirley Bay have been adjusted to represent the results that an azimuth-elevation mounted antenna would have obtained.

The change of polarization angle with time was due to two effects: the Faraday rotation and a smaller rotation arising from the geometry. This latter rotation was a consequence of the fact that the local vertical and horizontal directions were different for the widely separated sites. The geometric change in polarization angle for a particular day is shown in Fig. 5a as relative dipole rotation.

The variation of polarization angle with time can be computed approximately for quiet days by taking into account the expected variations in ionization density in the Earth's ionosphere. Such a computed curve is compared with the measured data in Fig. 5a. The best fit is used as the criterion to resolve the 180° ambiguity in the measured data. The assumed variation of electron density with height is shown in Fig. 5b. For each site the peak value of this assumed electron density was adjusted continually throughout the day to match the density inferred from the Central Radio Propagation Laboratory's 3-month prediction of "F-2 ZERO MUF".* The polarization angle computations were made using the exact expression for Faraday rotation and a dipole field to represent the Earth's magnetic field. The agreement between the computed and measured curves was usually good.

While the electron density distribution shown in Fig. 5b is not necessarily the actual distribution, it was found to produce reasonable agreement between the measured and predicted polarization angles. For this experiment the predicted polarization angle depends strongly on the electron density at altitudes near its peak value.

The power and polarization-angle plots shown in Figs. 3, 4 and 5a are made from 6-min averages of signal level and hence do not show short-period changes of the order of seconds. The total power has a rapid fading rate which can be deduced in a qualitative fashion from the signal samples shown in Fig. 2. The apparent polarization angle, as determined from the ratio of the two cross-polarized signal levels taken at one second intervals, also has a fine structure. These apparent fluctuations of polarization angle are about $\pm 10°$ in amplitude, with periods of the order of a few seconds. The period depends to a certain extent upon the Moon libration rate, and for this reason, it is perhaps an indication of the presence of a small depolarized component that is introduced at the Moon's surface.

There is one result that is not readily discernible from Figs. 4 and 5a but which is obvious from polarization plots made only from the ratio of the amplitudes of the two signals and consequently restricted to the first quadrant. There were times when the Faraday rotation was such that the polarization of the major component of the received signal was aligned with one of the two antenna polarizations. The channel where the polarization was at 90° to the major component of the incoming wave gave a signal 12–18 dB below the level in the channel oriented with the incoming wave. Thus the magnitude of the computed polarization angle went only to about 80°, although it was obvious that the polarization angle indeed went through 90°. Secondly, it was observed that in this condition there was little correlation between the detected signals in the two orthogonal channels. The signal in the low-level channel was always too large to be attributed to the mutual coupling of the two channels. At a polarization angle orientation of roughly 45° to both receiver feeds there was of course a high degree of correlation. The conclusion was that there existed a depolarized component of the signal reflected from the Moon some 15–20 dB below the major component. In addition, the fading rate of the weak channel was greater when the polarization angle was near 90° and this would lead one to believe that a larger portion of the Moon was responsible for the reflection of the depolarized component than was responsible for the main reflection.

The fading rate of the channel with the higher amplitude level was measured continuously

* *Basic Radio Propagation Predictions, CRPL Series D*, is issued monthly three months in advance by the National Bureau of Standards of the U.S. Department of Commerce.

during the experiment. The fading rate meter measured the number of positive crossings of the average signal value per unit time. The time constants used were a compromise between running time and freedom from effects of noise in raising the measured fading rate. Measured and predicted fading rates for a typical day are shown in Fig. 6. The top set of data was obtained from the signal received at Millstone, while the bottom shows similar results from Shirley Bay. The computed curve is based upon the predicted apparent libration data, and is 0·3 of the difference between the Doppler frequency of a signal reflected from the center of the lunar disc and that of a signal reflected from that portion of the limb having the maximum change in Doppler frequency due to the apparent Moon libration. The factor of 0·3 was found to give the best fit between the measured and predicted data. The agreement between the measured and predicted fading rate indicated that at least under quiet conditions the apparent Moon libration is responsible for the rapid signal fading. This dependence of fading rate on Moon libration has been reported by others.[2, 8, 10–12]

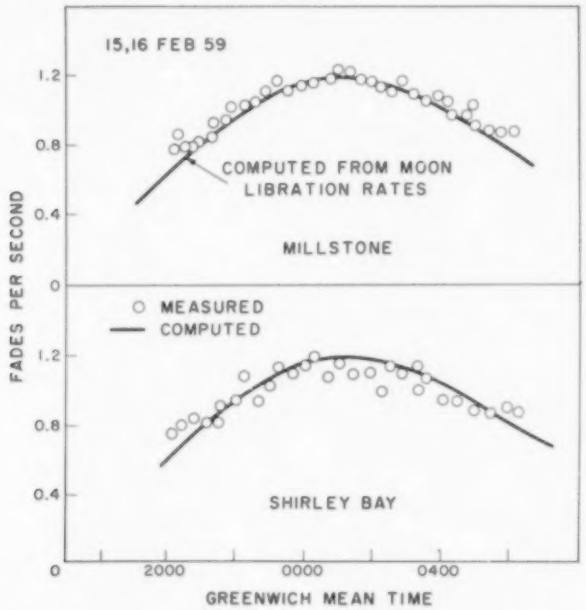


Fig. 6. Measured and computed fading rates for both receiving sites for a typical lunar day.

The fading rate is a fairly rough but rather easily obtained measure of the spectral bandwidth of the lunar signal. Actually knowledge of the power spectrum would allow one to deduce much of the reflection law of the Moon's surface. Because of the libration effects different portions of the Moon's surface reflect energy with different Doppler shifts. The power spectrum would give an estimate of the portions of the Moon's surface giving rise to reflection at a given Doppler shift and thus determine the extent of the Moon's surface effective in the reflection mechanism.

Power density spectra of the received signal were made from magnetic tape recordings of the i.f. signal. Two samples of such spectra are shown in Fig. 7, including a plot of the resolving power of the spectrum analyser. The top curve represents the Doppler spectrum for a quiet ionosphere. The bottom curve represents a period of auroral activity and will be discussed

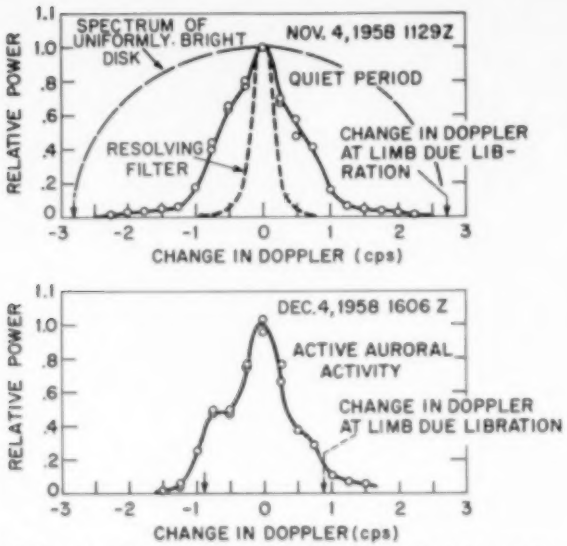


Fig. 7. Spectra of lunar signals obtained by making 1-min integrations of single-channel power passing through a $\frac{3}{10}$ c/s resolving filter. The Doppler change due to the change in the distance from the Moon to the observers during the 1-min interval has been removed.

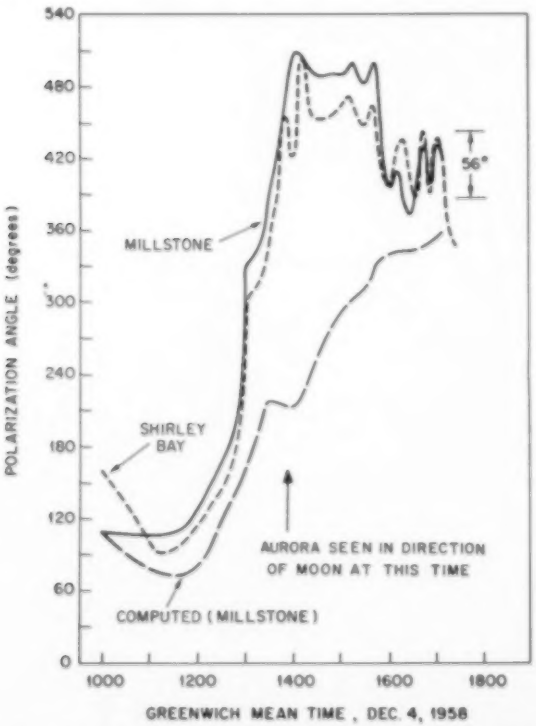


Fig. 8. The variation of the polarization angles at the two receiving sites for a day when a visual aurora was seen in the path to the Moon at College. The computed curve was based on a quiet ionosphere and computed as in Fig. 5a.

later. If the Moon's disc reflected uniformly the spectrum would be as shown in Fig. 7. From the appearance of the actual spectrum it can be seen that most of the received energy was reflected from the center portion of the Moon's disc. Incidental FM in the system, primarily in the tape recorder, may make the indicated spectrum wider than the actual spectrum of the signal. A central cap of approximately 0·2 fractional radius probably contains most of the effective reflection area. Another order of magnitude of frequency resolution would be required to determine whether the spectrum is essentially continuous or is composed of discrete frequency components.

During auroral activity several changes were observed in the character of the signal. The observed polarization angle exhibited rapid fluctuations, as shown in Figs. 8 and 9 for December 4, 1958, and February 16, 1959, respectively. The computed curves of polarization angle shown in these figures are based on a quiet ionosphere and consequently the agreement with the measured data was very poor. The presence of aurora along the line of sight to the Moon on December 4 was verified by

Fig. 9. The variation of the polarization angles at the two receiving sites for a day when radar backscatter from an auroral disturbance was observed in the path to the Moon. The computed curve was based on a quiet ionosphere and computed as in Fig. 5a.

visual and photographic observations at about 1400 hr as noted on Fig. 8. No radar return was observed from the aurora in this direction, probably because of the large angle between the line of sight and the normal to the magnetic field. Visual observations after about 1430 hr were obstructed by a partially overcast sky. Fig. 9 shows the effect of an auroral disturbance verified by the presence of 398 Mc/s radar back-scatter along the path to the Moon. At 2100 GMT, February 16, the radar display simultaneously contained Moon echoes and auroral

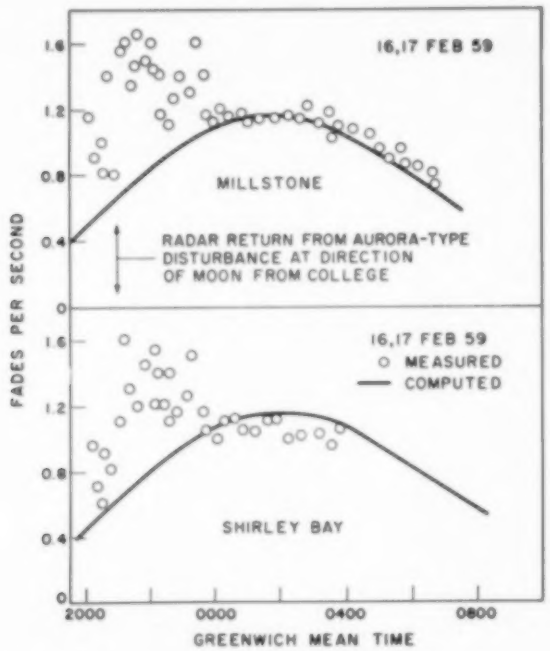


Fig. 10. Fading rates of the received lunar signal during times when the ionosphere was aurorally disturbed.

back-scatter. These observations were made during daylight hours at the transmitting site at College.

The large sudden fluctuations in polarization angle shown in Figs. 8 and 9 imply certain changes in the number of electrons in the ionosphere. To obtain values for the changes in electron density one must assume the location of the disturbance and the direction and magnitude of the magnetic field at this location. On the basis of a stationary magnetic field the series of fluctuations shown in Fig. 8 at 1630 hours and the fluctuation in Fig. 9 at 2100 hours represent an incremental change in integrated electron density along a vertical path in the Earth's ionosphere of at least 10^{17} electrons/m^2, assuming that the fluctuations in ionization occur below heights of 1000 km. It is reasonable to attribute the rapid change in polarization angle to a change in the electron density since very large changes in the direction of the magnetic field are required to explain the observed rapid fluctuations in polarization angle on this basis alone. For example, a change in the direction of the field of about 20° over a 400 km height interval in the ionosphere would be required to produce the 56° change in polarization angle observed on December 4. This is much larger than any known change in magnetic field direction, so it was concluded that these rapid fluctuations in polarization angle were due to rapid changes in ionization density. The rapid 75° fluctuation shown in Fig. 9 probably is due to a change in ionization

density. For this case also a large change in magnetic field direction is required to explain the fluctuation, if normal ionization densities are assumed.

The presence of auroral activity in the path to the Moon also had the effect of increasing the fading rate and hence in widening the spectrum of the received signal. This phenomenon is illustrated in Figs. 10 and 11, which show computed and measured fading rates for 2 days when auroral activity was found in the transmission path to the Moon. During the time of the auroral disturbance in the path the fading rates were appreciably larger than the computed values and rather erratic in nature. The polarization angles during the disturbed periods of Figs. 10 and 11 exhibited the rapid changes described previously. The fading rates for

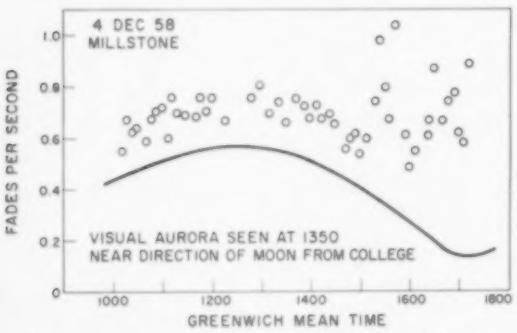


Fig. 11. Fading rates of the received lunar signal during times when the ionosphere was aurorally disturbed.

Millstone and Shirley Bay observed on February 16 are shown in Fig. 10. The increased rate at about 2100 hr coincides with the time that a radar return was observed in the direction of the Moon. Fig. 11 is a plot of the measured and computed fading rates at Millstone for December 4. The fading rate on this day was larger than predicted for the entire run and included the time when visual aurora was seen in the Moon's direction.

Another manifestation of the increased fading rate during aurorally active periods is the increased spread of the power spectrum. An example of this increase is shown by the lower curve in Fig. 7. This shows the spectrum of the signal in one channel on December 4 when aurora was present. The bandwidth is greater than would be produced by the libration effects of a uniformly illuminated moon. Thus the effects of modulation of the signal in a single linearly polarized channel by the aurorally excited ionosphere can mask the effect of the Moon's libration.

CONCLUSIONS

The results of this experiment indicate some of the important effects that the ionosphere will have on the propagation of an ultra high frequency signal into space. In particular the effects of an aurorally disturbed ionosphere are indicated. This experiment also served to confirm and extend the results of others in regard to Moon-reflection characteristics.

On several occasions when the propagated signal passed directly through an auroral disturbance there was no measurable decrease in received power. If there had been as much as 2 dB of signal attenuation for a time which is long compared with the rapid fading period, this most certainly would have been observed.

The measured frequency shift between the transmitted and received signals was found to be equal to the Doppler frequency predicted by the rate of change of the observer–Moon distance. Even during times of Moonrise and Moonset, and during auroral activity there was no discernible difference between the measured and predicted Doppler shifts. The extent of the Doppler frequency variation throughout the lunar day depends upon the transmitted frequency, the locations of the stations on the Earth, and upon the position of the Moon in its orbit. In the present experiment the Doppler shift changed approximately 1100 c/s throughout a lunar day.

The received signal had a power density spectral width of several cycles per second. The spectral width as measured on a single linearly-polarized channel was 2 or 3 c/s larger when the signal traversed an auroral disturbance. It has been shown that the fading rate is ordinarily proportional to the effective libration rate of the Moon. However this proportionality between fading rate and libration rate does not

hold at all times, since the fading rate is appreciably larger than expected when the signal traverses an auroral disturbance. The frequency of radio star scintillation is also increased when the star is viewed through auroral activity.[15] The two phenomena are quite possibly the result of the same type of signal perturbation even though the lunar signal is disturbed on the way to the Moon and is not disturbed on its return to Earth. The increased fading rate may be due in part to rapid polarization-angle fluctuations but is not explainable completely on this basis because the increase in fading rate observed in a single channel does not appear to depend upon the division of power between the two orthogonal channels. A similar experiment at a lower frequency would help to resolve this problem, as the fading rate due to libration would be smaller but fading due to the Faraday effect would be more rapid.

The use of circular polarization would eliminate any fading due to the rotation of the plane of polarization. This would also eliminate the possible requirement of following the plane of polarization.

It was found by using a cross-polarized receiving antenna that the plane of polarization of the lunar signal underwent a continuous change with time. Approximate curves representing the average total rotation as a function of time of day can be computed for days when the ionosphere is undisturbed. For aurorally disturbed days however there was very little resemblance between the computed and measured polarization curves. Large changes in polarization angle occurring over a period of a few minutes are assumed to be due to auroral ionization. These rapid changes sometimes represent an electron change of about 10^{17} electrons throughout a vertical column of 1 m^2 cross section. The electron density change depends upon the vertical extent of the additional ionization created by the auroral activity. For example, if the vertical extent were 10 km, the change in electron density would be of the order of 10^7 electrons cm^3.

A further point regarding effects which arise as a result of Faraday rotation should be emphasized; namely, the intensity of these phenomena is proportional to the square of the wavelength, so that at higher frequencies some effects reported here become less important.

The polarization angle was determined by assuming the incoming energy to be plane polarized and comparing the amplitudes of the signals induced in each of the two cross-polarized antennas. This measurement on a small time scale leads to apparent fluctuations of about 20° with periods depending to a certain extent upon Moon libration rate. It can be argued that this apparent small period fluctuation is not a real change in Faraday rotation because, if it were, these fluctuations would be more than 90° at 120 Mc/s, and fluctuations of this magnitude have not been reported.[8] It seems quite likely that multiple reflections somewhere on the Moon's surface may be the cause of this phenomenon. An experiment performed with a transmitter placed on the Moon's surface would resolve this problem.

An obvious application of the results of this and other lunar experiments is the use of the Moon as a reflector for communication circuits. This has been demonstrated during previous lunar experiments,[7] and as a continuation of the present experiment was demonstrated at 440 Mc/s by a test between the Millstone site and the Prince Albert Radar Laboratory in Saskatchewan, Canada. A recording of a message from President Eisenhower to Prime Minister Diefenbaker was used to modulate the 50-kW transmitter at Millstone. The modulation was single sideband with suppressed carrier and the lunar reflected signal was received with the 84-ft dish at Prince Albert. The message was received with a mean signal-to-noise ratio of about 15 dB using a 2 kc/s receiving channel and there was no serious quality deterioration due to fading.

In summary three results of this experiment that are of importance to Earth–space communication at ultra high frequencies should be emphasized. First, a system designed for space communication must have an antenna arrangement that avoids possible signal nulls due to Faraday rotation. If the radio energy is likely to

traverse auroral disturbances, rapid changes in polarization angle should be expected. Second, there should be negligible signal attenuation at u.h.f. due to auroral ionization. Third, the presence of auroral activity in the propagation path should increase both the rate and amplitude of the fading.

Acknowledgments—The following people were also actively engaged in and responsible for the performance of this experiment: L. E. Bird, S. J. Fricker, H. H. Hoover, W. C. Mason, H. G. Root and D. W. Swift of the Lincoln Laboratory; B. C. Blevis, J. W. B. Day and G. E. K. Lockwood of the Defence Research Telecommunications Establishment; and M. R. Berg, L. Kain, R. I. Presnell and J. C. Schlobohm of the Stanford Research Institute. The authors also thank the many other people who helped perform the experiment and in particular they acknowledge the co-operation of the personnel of the Geophysical Institute, University of Alaska, of the Coast and Geodetic Survey Laboratory at College, Alaska, and of the Millstone Hill Radar Observatory of the Lincoln Laboratory.

REFERENCES

1. J. Mofenson, *Electronics* **19,** 92 (1946).
2. F. J. Kerr and C. A. Shain, *Proc. Inst. Radio Engrs.* **39,** 230 (1951).
3. W. A. S. Murray and J. K. Hargreaves, *Nature, Lond.* **173,** 944 (1954).
4. J. H. De Witt, Jr. and E. K. Stodola, *Proc. Inst. Radio Engrs.* **37,** 229 (1949).
5. P. G. Sulzer, C. F. Montgomery and I. H. Gerks, *Proc. Inst. Radio Engrs.* **40,** 361 (1952).
6. F. B. Daniels and S. J. Bauer, *Nature, Lond.* **181,** 1392 (1958).
7. J. H. Trexler, *Proc. Inst. Radio Engrs.* **46,** 286 (1958).
8. I. C. Browne, J. V. Evans, J. K. Hargreaves and W. A. S. Murray, *Proc. Phys. Soc., Lond.* **B69,** 901 (1956).
9. B. S. Yaplee, R. H. Bruton, K. J. Craig and N. G. Roman, *Proc. Inst. Radio Engrs.* **46,** 293 (1958).
10. B. C. Blevis, *Nature, Lond.* **180,** 138 (1957).
11. S. J. Fricker, R. P. Ingalls, W. C. Mason, and M. L. Stone, *Nature, Lond.* **182,** 1438 (1958).
12. S. J. Fricker, R. P. Ingalls, W. C. Mason, M. L. Stone and D. W. Swift, *Computation and Measurement of the Fading Rate of Moon-Reflected UHF Signals, J. Res. Nat. Bur. Stand.* **64D,** 455, (1960).
13. J. V. Evans, S. Evans and J. H. Thomson, *The Rapid Fading of Moon Echoes at 100 Mcps*, p. 8, in *Paris Symposium on Radio Astronomy* (edited by R. N. Bracewell), Stanford University Press (1959).
14. J. S. Hey and V. A. Hughes, *Radar Observations of the Moon at 10 CM Wavelength, ibid.* p. 13.
15. C. G. Little and A. Maxwell, *J. Atmos. Terr. Phys.* **2,** 356 (1952).

ANALYSIS AND PREDICTION OF RADIO SIGNAL INTERFERENCE EFFECTS DUE TO IONIZED LAYER AROUND A RE-ENTRY VEHICLE

WILLIAM C. TAYLOR
Lockheed Missiles and Space Division

Abstract—Principally this report treats the problem of transmission of electromagnetic waves through the thermally ionized layer of gases in the shock layer around a hypersonic missile. Giving simplified theory, methods and graphic aids for approximate predictions of the problem, it also presents some specific results for a typical hemisphere-cylinder vehicle shape. While details of the theory allowing computation of temperature and the concentration of electrons and neutral particles are not given, a method which has been used is outlined and referenced.

INTRODUCTION

Background and History

It has long been known that weakly ionized gases have significant effects upon electromagnetic propagation. Especially important have been the experiences and analysis connected with ionospheric propagation and reflection. It was not surprising then that difficulties in telemetering have been encountered with the ionization produced because of shock-heating of the air about a missile re-entering the atmosphere with great velocities.[1–5]

Varying degrees of success have been realized by various investigators in attempts to predict theoretically the quantitative effects of the ion sheath upon a transmitted signal.[6–10] Experimental data, i.e. signal strength data actually received from such transmitters during re-entry, are usable in analysing the problem to some extent; however, there are several phenomena which may impair signal reception from the vehicle, and this fact makes the discernment of the various contributions difficult at best. There is every reason to believe, however, that thermal ionization of air is one of the most important of these causes, perhaps the sole serious cause in many cases.

Basic Nature of Interaction of EM Waves with Ionized Gases

When a weakly ionized gas is subjected to rapidly alternating electromagnetic fields, the chief phenomenon of interest is the motion of the free electrons, by far the lightest of the particles of significant concentration. In the absence of any neutral particles, the electrons will oscillate under the influence of the electric fields, thereby both receiving from and re-radiating energy to these fields. The convection current formed by this electronic motion is in opposition to the normal displacement current associated with free space propagation, and for sufficient electron concentration, will prohibit completely any propagation much as do waveguides for low frequencies. The electron cloud may then be thought of as rejecting any attempt to propagate through it even though no energy is absorbed in the process. The presence of neutral particles in the paths of the electrons will result in random collisions, disturbing the orderly re-radiation process; the result is a net loss of energy by the wave to the ionized gas.* Since both electrons and neutral particles constitute the environment about a re-entering missile, both of these mechanisms, absorption and rejection, play important roles in affecting the transmission from the vehicle.

* In a weakly ionized gas, i.e. a gas with less than one per cent of the total particles ionized, electron–electron collisions are negligible, and the lack of mobility of the positive ions compared with electrons minimizes their effects.

PROCEDURES AND THEORY

Orientation

The procedure for computing electromagnetic wave propagation through the ionized sheath can be logically divided into three phases:

(1) Calculate the concentration of electrons and of massive particles (molecules, atoms and positive ions) which have appreciable concentrations in the region about the missile. These concentrations must be determined along with temperature throughout the re-entry phase of missile flight.

(2) Calculate the macroscopic electrical properties of the sheath using the data from phase (1) and the appropriate "conductivity theory" for the ionized air. This involves calculation of the average electron collision frequency and requires knowledge of the wave frequency and data from other sources on the electron collision cross sections of the massive particles. The end result is the determination of the electromagnetic propagation constant throughout the ionized region for the re-entry portion of the trajectory.

(3) Calculate the propagation through the shock layer. In this final phase, the effects of reflection of the wave from the shock layer and attenuation of the wave in the shock layer are included. An approximate solution is necessary since the complicated geometry of the vehicle and the ion sheath make the exact solution intractable.

These three phases are discussed consecutively in the following sections.

Determination and Particle Densities

The particle densities of the various species can be obtained from the results of Ref. (11),

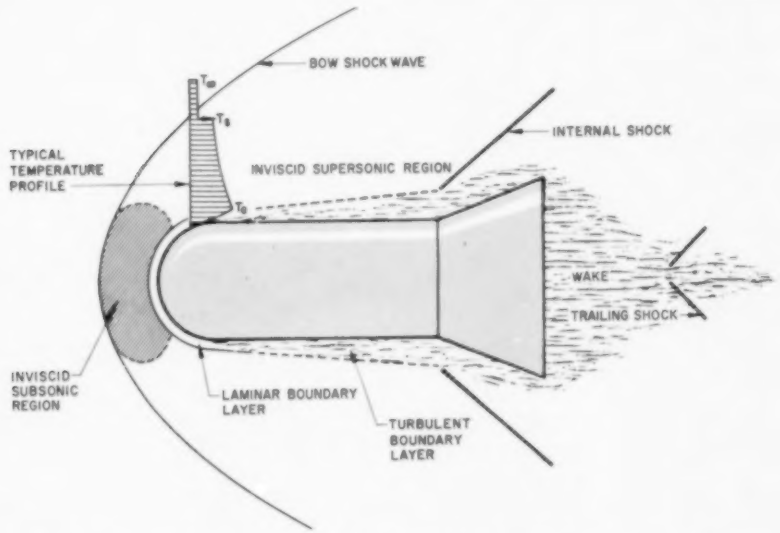


Fig. 1. Schematic representation of flow field around body in hypersonic flight.

once the thermodynamic state of the gas is known at each point in the flow field. Determination of the state of the gas requires knowledge of the entire flow pattern. Thus, in order to compute the state point of the gas, it is first necessary to solve the fluid-dynamic problem of the flow about a body in hypersonic flight.

The new field can be logically divided into a number of separate regions, as indicated by the mathematical character of the governing equations. These regions are depicted schematically in Fig. 1. In the inviscid subsonic region, the governing flow equations are elliptic, and hence numerical solutions are faced with stability

problems. Only recently have numerical techniques been developed,[12] which permit accurate determination of the flow in the inviscid subsonic region. At the time of preparation of this report, however, none of the numerical techniques accounted for real gas effects, and as a result they could not be used for this study.

high-temperature gases are confined to the "shock layer", i.e. the region between the bow shock wave and the outer edge of the boundary layer. Fortunately, it is possible to predict the state of the gas on the two surfaces of revolution which bound the shock-layer region.

For computations leading to this report, blast

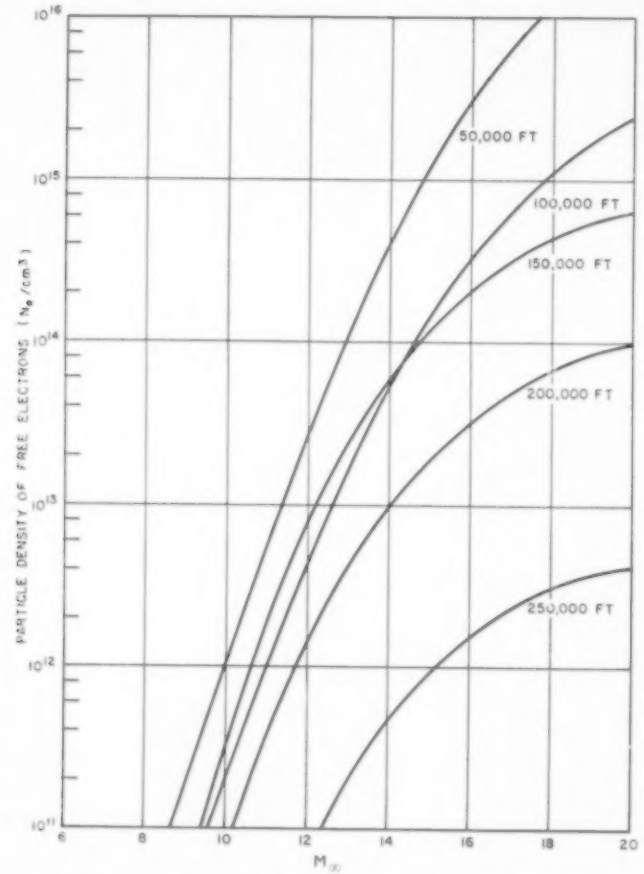


Fig. 2. Electron density at stagnation point vs. Mach number.

The flow in the inviscid supersonic region can be computed by the method of characteristics when sufficient initial data are given. Unfortunately, the initial data must be obtained from a solution for the subsonic region, and as noted above, solutions for the subsonic region could not be obtained. Furthermore, a characteristics program which included real gas effects was unavailable.

Referring again to Fig. 1, it is seen that the wave theory[13] was used to predict the shape of the bow shock wave, and the charts of Ref. (14) were used to determine the state of the gas immediately behind the shock. Conditions at the outer edge of the boundary layer were determined, using measured pressure distributions and the assumption of isentropic flow. With the state point of the gas determined on these surfaces, the results of Ref. (11) were used to compute particle densities. Fig. 2 shows electron

density at the stagnation point of any missile as a function of Mach number at several altitudes. To obtain particle densities in the region between the shock and the edge of the boundary layer, it was assumed that the temperature and density varied linearly in a radial direction through the shock layer.

The laminar and turbulent boundary layers which develop adjacent to the surface are amenable to analysis. However, they can generally be dismissed as an important source of ions because of the fact that the temperature drops rapidly in these thin layers, and the electron concentration is an exponential function of temperature (see Fig. 3).

Only if the surface temperature exceeds about 2500°K, will significant ion concentrations exist next to the skin due to ionization of air. Another situation in which significant concentration of ions would exist occurs when an ablating material is used for heat dissipation. Under these conditions, the ionized plasma would not be separated from the body surface by relatively cool boundary layer gases. The quantitative results detailed in this report account for air ionization effects alone.

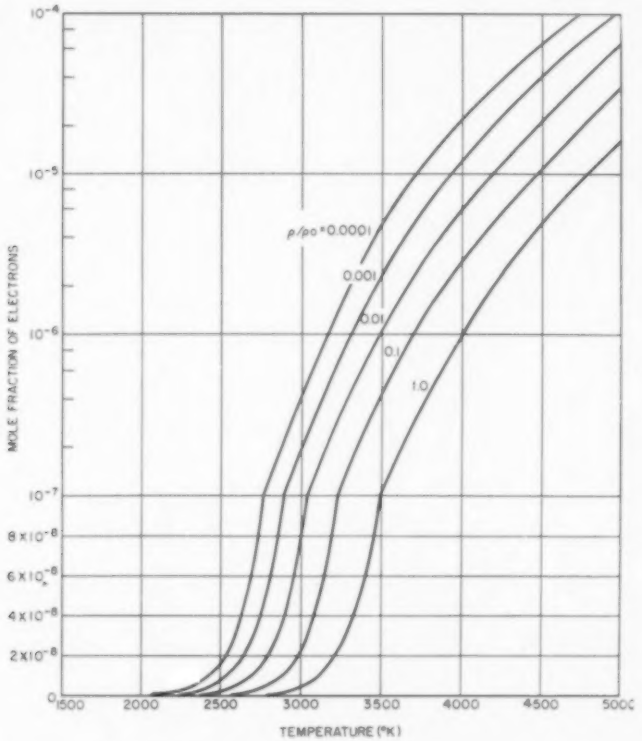


Fig. 3. The mole fraction of electrons in air (ratio of N to total number of particles per unit volume) as a function of temperature for 5 densities. ρ_0 is standard atmospheric density.

The final flow region depicted in Fig. 1 is the wake. At the present time, there is insufficient theoretical and experimental basis for predicting the flow in the wake.

In summation, methods exist for predicting most of the flow around blunt bodies in hypersonic flight. However, at the time this report was being prepared, these exact methods were either in the process of development or were being modified to incorporate real gas effects.

Plasma–Wave Interaction and Plasma Properties

Propagation Constant

The propagation constant $\gamma = \alpha + j\beta$ is used to describe the attenuation and phase of a wave in traversal through a medium. The propagation factor is thus

$$e^{-\gamma z} = e^{-\alpha z} e^{-j\beta z}$$

where z is the direction of propagation in the infinitesimal region described by α and β. In terms of the conductivity σ, the permittivity ϵ and the permeability μ of the plasma, the propagation constants for a plane wave of radian frequency ω are given by

$$\alpha = \omega \sqrt{\left\{\frac{\mu\epsilon}{2}\left[\sqrt{\left(1 + \frac{\sigma^2}{\omega^2\epsilon^2}\right)} - 1\right]\right\}} \quad (1)$$

and

$$\beta = \omega \sqrt{\left\{\frac{\mu\epsilon}{2}\left[\sqrt{\left(1 + \frac{\sigma^2}{\omega^2\epsilon^2}\right)} + 1\right]\right\}} \quad (2)$$

so that α and β are positive and real.

Lorentz Solution for σ and ϵ

The permeability of a weakly ionized plasma is generally considered that of free space, while σ and ϵ depend upon the electron concentration and upon the nature and frequency of the collisions of electrons with neutral particles. The collision process is, in turn, generally dependent upon the temperature, electric field strength and the density and collision cross sections of the massive particles. The Lorentz conductivity theory develops relatively simple quantitative relationships by solving the equation of motion of an electron in a gas under the influence of two types of forces, the electric field and collisions with neutral particles.[15] According to this theory, the convection current density is given by

$$i = E\left(\frac{Nq^2}{m}\frac{\nu}{\nu^2 + \omega^2} - j\frac{Nq^2}{m}\frac{\omega}{\nu^2 + \omega^2}\right) = \sigma^* E \quad (3)$$

where E is the electric field intensity, N is the electron concentration (number per m^3), q is the electronic charge in coulombs, and σ^* is the complex conductivity. (Unless otherwise mentioned, rationalized m.k.s. units are utilized in this report.) In an alternate representation, the imaginary part of the conductivity is made equivalent to a negative electric susceptibility, and the real part of the complex conductivity is retained. This gives

$$\epsilon = \epsilon_0\left[1 - \frac{Nq^2}{\epsilon_0 m(\nu^2 + \omega^2)}\right] = \epsilon_0\epsilon_r \quad (4)$$

and a real conductivity

$$\sigma = \frac{Nq^2}{m}\frac{\nu}{\nu^2 + \omega^2} \quad (5)$$

Substituting these values into equations (1) and (2) gives

$$\alpha = \frac{\omega}{\sqrt{2}c}\sqrt{\left\{-\epsilon_r + \sqrt{\left[\frac{\nu^2}{\omega^2}(1 - \epsilon_r)^2\right]}\right\}} \quad (6)$$

and

$$\beta = \frac{\omega}{\sqrt{2}c}\sqrt{\left\{+\epsilon_r + \sqrt{\left[\frac{\nu^2}{\omega^2}(1 - \epsilon_r)^2\right]}\right\}} \quad (7)$$

where c is the velocity of light. When the positive roots are chosen in the above forms, α and β will always be positive and real.

An ionized medium has a transmission line analog,[15] the equivalent circuit of which is shown in Fig. 4. The Lorentz conductivity theory

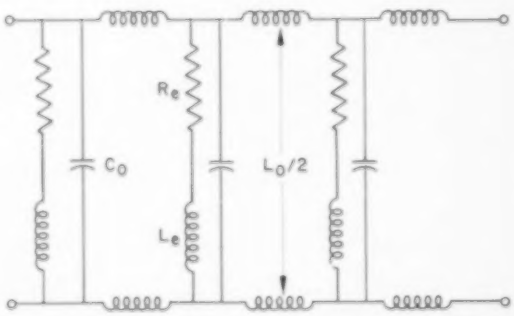


Fig. 4. The approximate equivalent circuit of a weakly ionized plasma.

gives $R_e = m\nu/Nq^2$ and $L_e = m/Nq^2$. If $\nu = 0$ and N is of such value as to satisfy the equation

$$\omega^2 = \frac{Nq^2}{\epsilon_0 m} = \omega_p^2 \quad (8)$$

the convection current through L_e is equal and opposite to the displacement current, and propagation ceases. As seen by equations (4)

and (7), this condition corresponds to $\epsilon_r = 0$ and, therefore, $\beta = 0$.

In fact, further examination shows that for negative as well as vanishing ϵ_r, propagation cannot occur when $\nu = 0$. For this reason, the value of ω given by equation (8) is called the "critical frequency" (sometimes "plasma" frequency) and is well known in ionospheric research. For finite ν, $\epsilon_r = 0$ when

$$(\omega^2 + \nu^2) = \frac{Nq^2}{\epsilon_0 m} \tag{8a}$$

but β remains finite for all ϵ_r. However, the attenuation constant becomes quite large for $\epsilon_r < 0$, and propagation is similar to that in metals, with field strengths attenuated to $1/e$ of the initial value within a "skin depth" $\delta = 1/\alpha$.

Equations (6) and (7) can be conveniently rewritten to give expressions for normalized α and β in terms of only two dimensionless parameters which are ratios of any two of the "frequencies", ω, ν and ω_p. Any of the following three pairs can be used

$$\frac{\nu}{\omega} \text{ and } \frac{\omega_p}{\omega}$$

$$\frac{\nu}{\omega_p} \text{ and } \frac{\omega}{\omega_p}$$

or

$$\frac{\omega}{\nu} \text{ and } \frac{\omega_p}{\nu}$$

The resulting expression is most useful for describing a situation where the normalizing frequency remains essentially constant, making

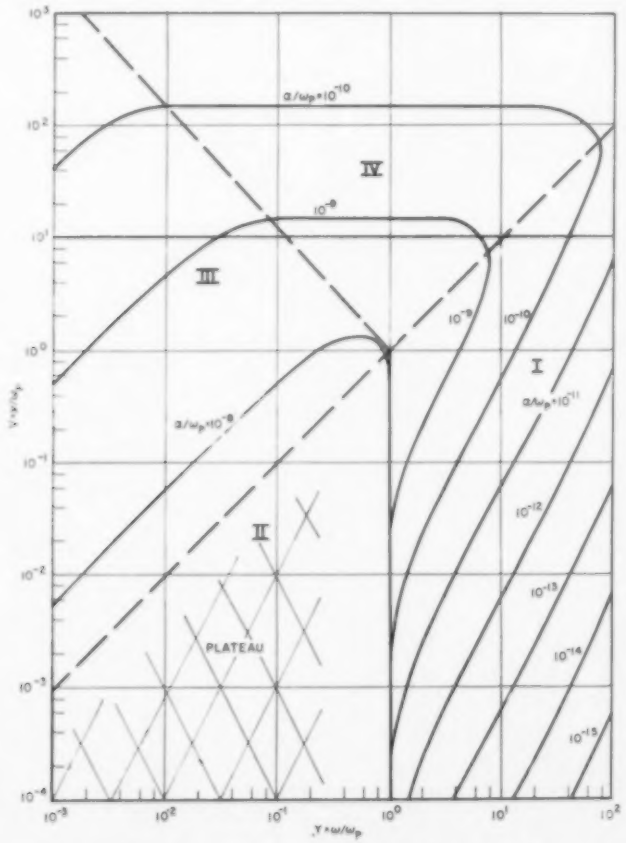


Fig. 5. Representation of the surface α/ω_p as a function of $V = \alpha/\omega_p$ and $Y = \omega/\omega_p$ by contour lines of constant α/ω_p.

detailed study of α and β somewhat simpler. A typical expression for α is

$$\frac{\alpha}{\omega_p} = \frac{Y}{\sqrt{2}c}\left\{-\left(1 - \frac{1}{Y^2 + V^2}\right) + \left[\left(1 - \frac{1}{Y^2 + V^2}\right)^2 + \frac{V^2}{Y^2}\left(1 - \frac{1}{Y^2 + V^2}\right)\right]^{1/2}\right\}^{1/2} \tag{9}$$

where $Y = \omega/\omega_p$ and $V = \nu/\omega_p$. The general behavior of α and β over a wide range of the variables can be well illustrated by graphical means. Naval Research Laboratory[16] has produced the several families of curves over a wide range corresponding to all three sets of the dimensionless parameters above. An alternative graphical presentation is that of a surface. Equation (9) is represented in this way in Fig. 5 (two-dimensionally) by contour lines of constant (normalized) attenuation. The dashed lines and the heavy dark line (along $\omega/\omega_p = 1$) divide this "surface" into four important regions of interest, all of which will be encountered during a typical ballistic re-entry. In Region I, attenuation decreases with increasing ω but increases with increasing collision frequency. Region II is characterized by an essentially constant, α/ω_p, independent of both ω and ν; in fact, the "plateau" of this region is analytically described by $\omega_p > \omega > \nu$, giving from equation (6),

$$\alpha \approx \omega_p/c$$

Region III exhibits a slow variation of α/ω_p just opposite to Region I, i.e. α/ω_p increases for increasing ω but decreases for increasing ν.

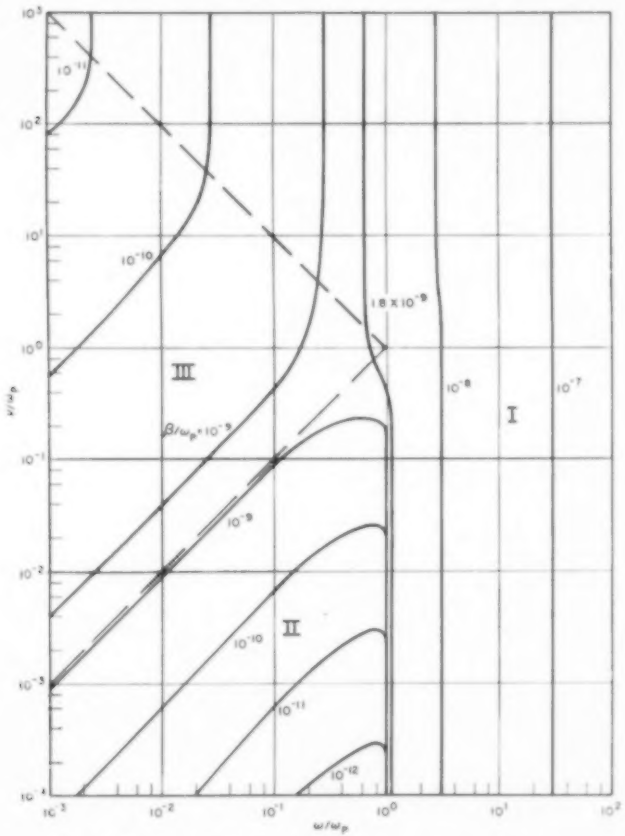


Fig. 6. Contour lines of constant β/ω_p vs. ν/ω_p and ω/ω_p.

Finally, Region IV shows α/ω_p independent of ω and varying inversely with ν. The most interesting aspect of the surface is the extreme slope at $\omega/\omega_p \approx 1$ where $\nu/\omega_p < 1$, giving rise to the term "critical" frequency. It is due to this "effect" that an increase in ω can effectively solve the transmission problem in higher altitudes (where $\nu/\omega_p < 1$) if ω can exceed ω_p. In lower altitudes, where ν may exceed ω_p even though ionization is still high, the attenuation can be reduced if $\omega^2 \gg \nu^2$. The behavior in Region III is the basis for the proposal in some extreme cases to reduce ω in order to reduce attenuation; obviously such a measure cannot make as drastic a reduction per change in ω as in the case of an increase in ω which takes from Region II to I, but it may be the only recourse for some instances. Ideally, it might be desirable to place a low-frequency transmitter at the stagnation point (where N is a maximum) and a high-frequency transmitter along the afterbody. When the latter signal is cutoff, the low-frequency transmitter has the best chance of transmission, the high ω_p taking it far to the left in Region III. More quantitative inspection shows that a ratio as great as 10^3 between the ω's may be required for the low frequency to be advantageous.

The qualitative variation of α with ω_p, as well as with ν and ω, can be deduced from Fig. 5. Going from upper right to lower left along any straight line of unity slope (along which $\nu/\omega =$ constant but ω_p increases), it is seen that α/ω_p will always increase or remain constant which means that α itself will increase rapidly and monotonically with ω_p in all cases. It is for this reason that N, hence temperature, is always quite significant in determination of plasma–wave interaction.

Because β becomes much less than α in the situations of greatest interest, it seems to be of less importance in this survey of plasma properties. However, because of its role in reflection, its general behavior as a function of ω, ω_p and ν should be noted, and is shown in Fig. 6, analogous to Fig. 5. Region I, including the entire surface where $\omega/\omega_p > 1$, is characterized by β/ω_p varying linearly with ω/ω_p independent of ν/ω_p; this corresponds to a constant value of β, $\beta_0 = 2\pi/\lambda_0$, the free-space phase constant. Region II indicates rapid variation in β/ω_p with both co-ordinates, and the most important feature is a deep minimum where ω/ω_p is slightly less than unity. From this "trough", β/ω_p will rise rapidly with any change in ω/ω_p, this variation being linear for decreasing ω/ω_p. Also, β/ω_p shows an important linear (direct) variation with ν/ω_p in this region.

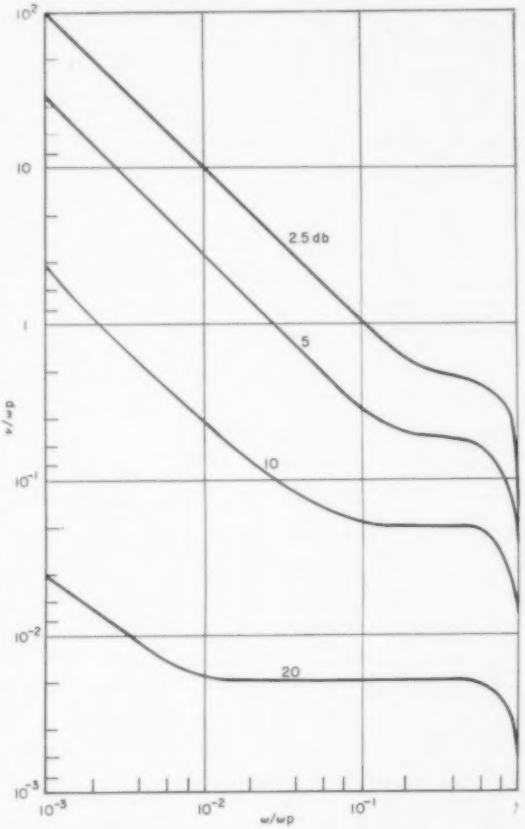


Fig. 7. Contour lines of constant attenuation (dB power) due to impedance mismatch at a boundary as a function of ν/ω_p and ω/ω_p.

The variation of β in Regions II and III is important in the computation of the reflection coefficient or, more important to this study, the transmission into a semi-infinite plasma at a discrete boundary. Fig. 7 shows lines of constant transmission coefficient as a function of ω/ω_p and ν/ω_p. The lines are identified by the values of power reduction (dB) effected by the mismatch

at this boundary. Note that no attenuation due to propagation over a finite distance is taken into account here.

Calculation of Collision Frequency

The collision frequency of an electron with homogeneously distributed particles depends upon its velocity and upon the effective size of the particles presented to the electron. Since a large number of electrons possess distributed rather than equal thermal velocities, and since the collision cross section is generally a function of velocity, an accurate collision frequency describing a macroscopic property will be an averaged one accounting for these facts. However, a useful approximate definition, which utilizes a constant collision cross section value for a given species, is derived from the product of the mean thermal velocity $\bar{v}$ of the electrons times the total cross section per unit volume of all the species present, i.e.

$$\bar{\nu} = \bar{v} \sum_i \rho_i Q_i \tag{10}$$

where

$$\bar{v} = \left(\frac{8kT}{\pi m}\right)^{1/2}$$

ρ_i is the number density of the i^{th} species
Q_i is the collision cross section of the i^{th} species
k is the Boltzmann's constant
T is the temperature (°K)

The Lorentz conductivity theory, which is the source of the quantitative results of this paper, inherently contains certain approximations not yet mentioned;[17, 18] however, the difference between the results obtained using the simpler Lorentz law and the more accurate theories[19] are usually found to be smaller than the probable errors encountered in other steps of the computations procedure outlined in this paper.

However, one influence exerted upon the collision process, which is an exception in some cases and has not been accounted for, is that of relatively high electric field intensities encountered because of nearness of the gas to the antenna. It is simplest to assume, as is done in the Lorentz case, that the energy imparted to an electron by the electric field is small compared with the thermal energy it possesses, i.e. that[18]

$$E^2 \ll \frac{\sigma k T \omega^2}{q^2} \frac{m^2}{M}$$

where M is the mass of the neutral particle in the collision; if this inequality holds, then ν, hence σ, is independent of E. For air at 3000°K, at 200 Mc/s, this means E must be much less than 4 V/cm. Sisco and Fiskin[20] have shown that the principal effect of stronger fields is, as one might expect, to increase electron velocities (hence collision frequencies) when the period of the wave is great enough. For the temperature and frequency noted above, with a mean free path of 0·15 cm, this effect begins around $E = 10^{-1}$ V/cm; from this point, increasing E by a factor of 10 will increase ν by a factor of about 3. The magnitude of the error involved in neglecting these effects is seen to vary widely, depending upon the radiator and its frequency as well as density and pressure. However, in the case when attenuation is high, the region of space where the above approximation might be poor is expected to be confined to a relatively thin layer near the antenna.

A limiting case for increase of E occurs with r.f. breakdown conditions, which, depending upon particle density and frequency of the radiator, are often met with common transmitters on ballistic missiles during both exit and re-entry.[21] If electron densities due to thermal ionization are below the critical concentration (such that $\omega_p < \omega$), the breakdown losses will usually exceed those determined in the preceding analysis. However, for $\omega_p > \omega$, the conditions which effect transmission losses in breakdown already exist to the degree that further ionization by the electric field will likely be masked in effect.

Analysis and Prediction of Total Attenuation

The propagation properties of the medium surrounding a hypersonic missile have been shown to vary all the way from those of metals to lossy dielectrics and free space. In the case of a plasma where an imaginary conductivity can play a role more important than the real σ, it

seems appropriate to define a good propagating medium as one with $\beta^2 \gg \alpha^2$ and a poor one with $\alpha^2 \gg \beta^2$. Because of the rapid variation of α and β in the region $\alpha \approx \beta$, there exists little middle ground in the plasmas between the cases of "good propagator" and "poor propagator". Illustrating this, Figs. 8 and 9 show approximate

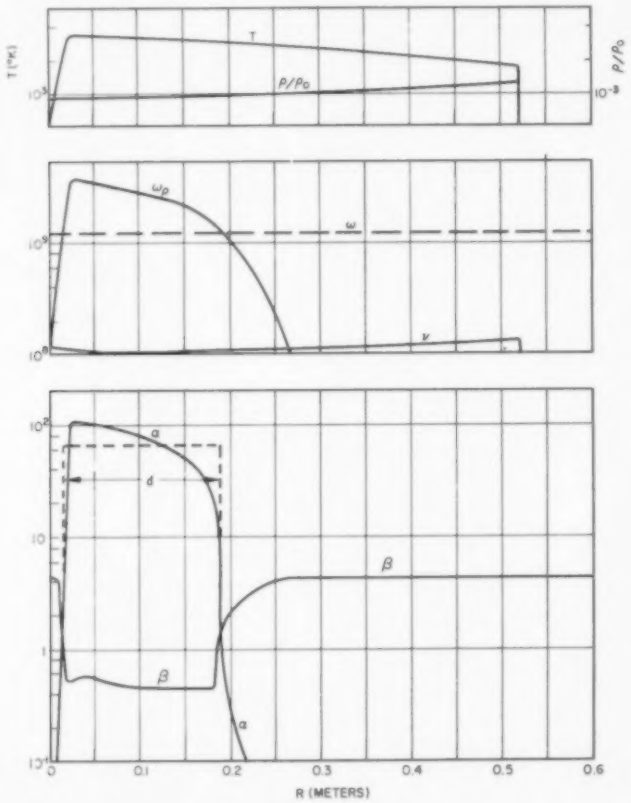


Fig. 8. A typical simultaneous radial variation (out from a blunt-nosed, cylindrical body) of plasma properties as computed or assumed at an altitude of approximately 200,000 ft and a velocity of about 16,000 ft/sec. A typical profile of computed or assumed plasma properties varying radially out from a blunt-nosed, cylindrical body. Altitude is approximately 200,000 ft and velocity about 16,000 ft/sec.

high altitude ($\omega^2 \gg \nu^2$) and low altitude ($\omega^2 \ll \nu^2$) spatial distributions of α and β corresponding to the exhibited temperature and density distributions. These correspond to computed or assumed values found varying radially outward from a relatively blunt-nosed, cylindrical body with a velocity of approximately 16,000 ft/sec in both cases. These and other computations lead to a "plasma model" which is illustrated in Fig. 10. The ion sheath is pictured as having a core of metal-like propagation properties with varying α, β and thickness d, with $\alpha \gg \beta$. This core is "coated" by a dielectric also with varying thickness and loss characteristics, with the thin boundary between core and coating corresponding to the region around $\alpha \approx \beta$. The "core" will extend back as far as temperatures are high enough that $\omega_p/\omega > 1$. Since temperatures are generally lower on a cylindrical after-body than a cone or similar shape, a cylinder is generally preferred for antenna location. It seems reasonable to expect that such a core does exist in an undetermined thickness in the shock layer since actual telemetered signals have experienced the great attenuation indicated in Refs. (1–4). It

also seems likely that the sharpness of the fade experienced in many cases corresponds to the quick transition in time into "cut-off" of the layer over the transmitting antenna.

The simplest quantitative evaluation of the total attenuation is accomplished by considering the plasma as a homogeneous flat sheet of infinite extent upon which a plane wave is (normally) incident. The thickness and propagation properties of this "equivalent sheet" are largely determined by those of the "metal-like" portion of the sheath, since its effects are considerably more important than a dielectric coating, the only significant thickness of which is on the side of the sheath opposite the transmitter. The boundaries of this "sheet", located where $\alpha = \beta$, are considered discrete because α and β change rapidly over a thickness which is very small compared with wavelengths considered here.

The radome-type problem of determining the power transmission coefficient through the sheet is simplified because the extreme attenuation upon a single traversal through the sheet will make the effect of multiple reflections negligible. The fraction of power transmitted into space beyond the sheet of that originally incident is then computed as the product of the surface mismatch transmission coefficients and the attenuation factor associated with propagation within the sheet, $\exp(-\alpha d)$, where d is the assumed thickness of the homogeneous sheet

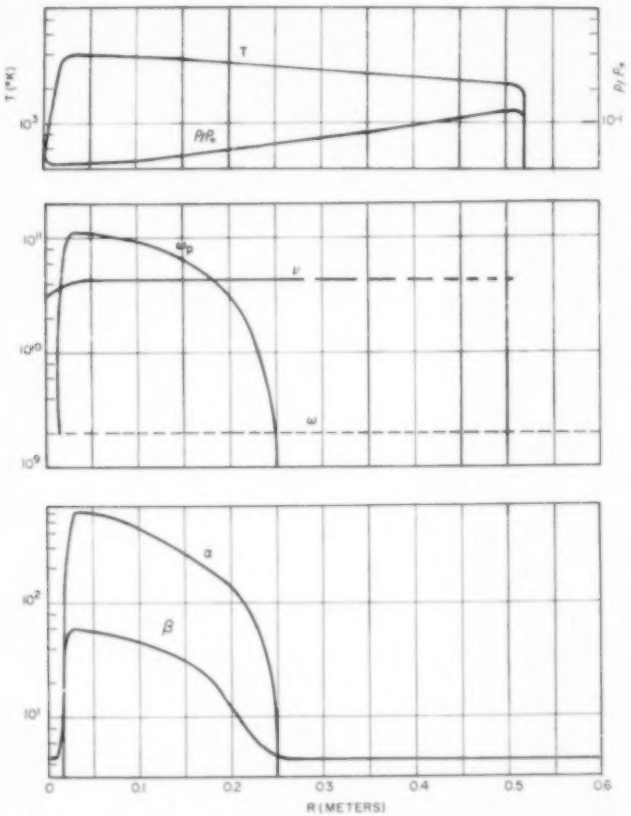


Fig. 9. A low-altitude profile of radial variation of plasma properties as computed or assumed for a blunt-nosed cylinder with a velocity or about 16,000 ft/sec.

Expressed in decibels of power attenuation, the total attenuation is then

$$A(\text{dB}) = 2 \cdot \log \frac{1}{|\tau|^2} + 8{\cdot}7\,\alpha\, d \qquad (11)$$

where

$$|\tau|^2 = 1 - |r|^2 = \frac{4n}{(n-1)^2 + k^2}\,|r|^2$$

is the power reflection coefficient for a single boundary, and the factor 2 accounts for the second boundary reflection. This concept of the problem, besides assuming the plane wave–plane sheet picture, neglects the history of the large amounts of energy reflected back toward the missile, including the effect of antenna detuning.

An approximate evaluation of d can easily be made if T is assumed to vary linearly from its maximum near the skin since ρ/ρ_0, which also varies slowly, is not so critical in determination of ω_p. The metal-like sheet is assumed to extend out to where T is of a low enough value that $\omega_p = \omega$. As Figs. 8 and 9 show, α varies within d enough that a selection of the constant α for the sheet should be somewhat less than the maximum value near the skin. The variation of α and β is expected to be more complex at other locations than along the cylindrical afterbody.

Langberg *et al.*[9] have rigorously solved the wave equations for an inhomogeneous medium to compute attenuation when the wavelength is considerably greater than the thickness of the plasma. The real and imaginary parts of the propagation constant in this case must be determined by numerical solutions of simultaneous differential equations. The attenuation can then be evaluated by integrating the attenuation function through the plasma and accounting for the change of the phase function across the plasma.

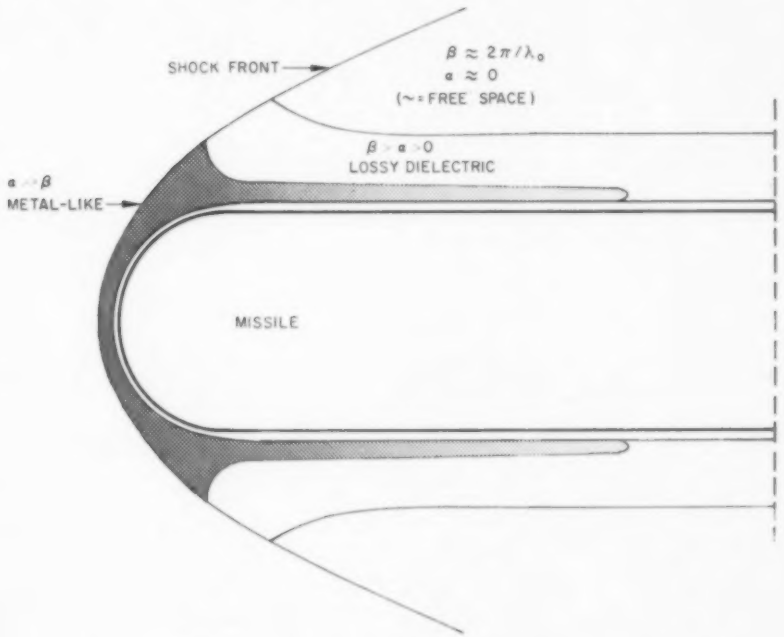


Fig. 10. Cross section of simple system considered similar in propagation properties to typical ion sheath during re-entry.

PRESENTATION AND DISCUSSION OF TRAJECTORY PREDICTIONS FOR A TYPICAL ANTENNA LOCATION

The most important extension of the examples given in previous sections of this paper is the

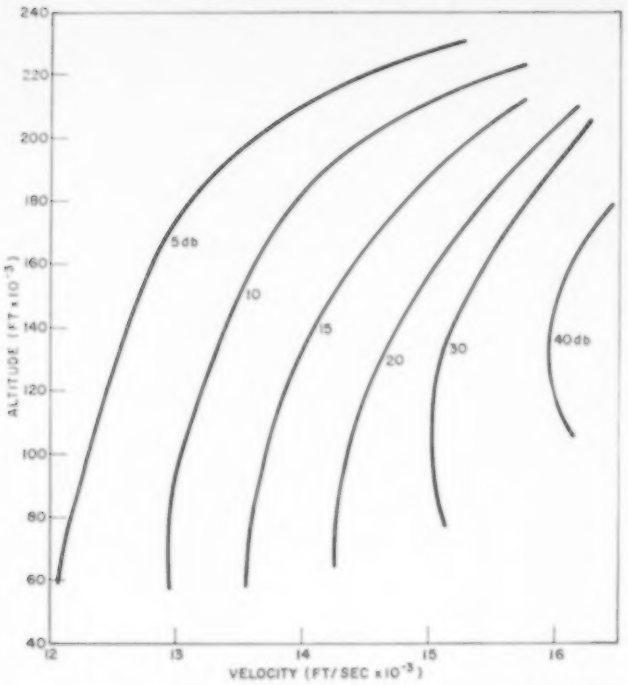


Fig. 11. Approximate contour lines of constant attenuation for a 200-Mc/s signal. The signal is radiated from a position well aft of the stagnation region on a hemispherically capped re-entry body. The lines are given as a function of both velocity and altitude.

construction of a velocity–altitude chart of total attenuation contours and the graphing, usually against time, of the deductions from this chart concerning attenuation on a given trajectory. Fig. 11 is such a contour chart for a 200-Mc/s signal radiated from a typical position on a cylindrical afterbody. Fig. 12 is a presentation of predictions concerning a typical re-entry trajectory as deduced from Fig. 11 upon the assumption that 20 to 30 dB of attenuation from this source is tolerable before the signal is lost by the monitoring stations.

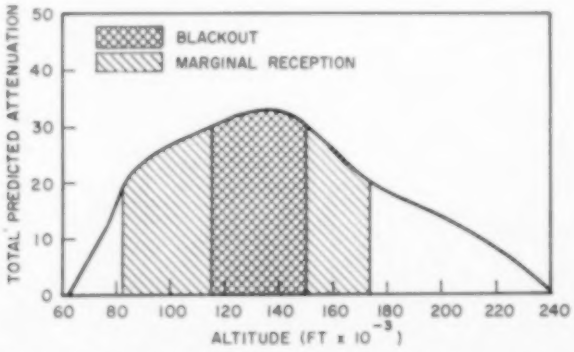


Fig. 12. Total predicted attenuation during a typical intermediate trajectory as a function of altitude. Deductions concerning reception allow 20–30 dB before complete loss of signal.

A chart similar to Fig. 11 can be constructed by first computing according to equation (11) the attenuation in a propagation "path" from the location considered, using several sets of values of altitude and velocity in the ranges of interest. After interpolation (usually graphical) between these results in order to determine the velocity attitude co-ordinates of 5 dB, 10 dB, etc. points, the points are then plotted on a grid such as Fig. 11. The drawing of highly approximate contour lines to fit the points smoothly is executed with as much an application of good judgment as real data when points are relatively scarce. However, the results seem generally sensible and useful, and have shown satisfactory agreement with received data from actual missile tests in a majority of cases.

The predicted effect of an increase of frequency to 10 kMc/s for the same position is that the plasma will never reach "cut-off" for velocities below 16,000 ft/sec; thus the total attenuation will apparently remain below 5 dB.

CONCLUSIONS

Experience and theory lead to the following conclusions:

(1) Signals in the 200-Mc/s region, radiated from typical antenna locations, will experience significant attenuation during a portion of the trajectory when the peak re-entry velocity of a ballistic missile exceeds about 11,000 or 12,000 ft/sec. Fig. 11 shows approximate predictions for an antenna placed well behind the stagnation area on a cylindrical body.

(2) An increase of frequency can increase, decrease or leave essentially unchanged the attenuating effect of the plasma on the signal, depending largely upon the electron concentration. Usually, for environments found around intermediate range (or shorter) missiles, increasing the frequency will decrease the attenuation at least in time duration, and for sufficiently high frequencies, may reduce it to a negligibly low level. For the case cited above, predictions for 10,000 Mc/s indicate that attenuation will be below 5 dB for velocities below about 16,000 ft/sec.

(3) For simple vehicle shapes, attenuation of a signal (200 Mc/s or greater in frequency) will usually decrease with distance of the radiator from the stagnation point since ionization intensity is less in regions where the vehicle surfaces are more parallel to direction of travel. Irregularities, such as flares, which produce internal shocks are exceptions.

REFERENCES

1. Army Ballistic Missile Agency, *Preliminary Report on the Re-entry of Missile AM-3 with Reference to RF Attenuation*, DG-TN-208 (Dec. 1957).
2. Army Ballistic Missile Agency, *Preliminary Report on the Re-entry of Jupiter Missile AM-5 as Compared to Missile AM-3 with Reference to Radio Frequency Attenuation*, DGI-TM-5-58 (2 Jun. 1958).
3. Army Ballistic Missile Agency, *Preliminary Evaluation of Jupiter AM-6A with Reference to Radio Frequency Attenuation*, DG-TM-150-58 (18 Aug. 1958).
4. B. M. Berkowitz, *Received Signal Strength Data, Missile 4B*, Gen. Electr. Co., MOSD, DIN 1014-03-2 (2 Sep. 1958).
5. E. F. Pischel, *Radio Signal Attenuation Induced by Re-entry Ionization*, Lockheed Missile Syst. Div. LMSD-3186, Sunnyvale, Calif. (15 Jul. 1957).
6. S. C. Lin *et al.* "*Hyperatmospheric and Hypersonic Phenomena; Properties of Missile Materials, Electrical Properties of High Temperature Air, and the Telemetering Problems of Long Range Ballistic Missiles*, in *Trans. First Tech. Symposium on Ballistic Missiles*, Vol. 4 (June 1956). See also *AVCO Res. Lab. Semi-annu. Rep. on Re-entry Body Problem* (1956–1958).
7. W. B. Sisco and J. M. Fiskin, *Electromagnetic Wave Propagation in Shock-Ionized Media*, Douglas Aircraft Co., Long Beach, Calif.
8. C. P. Sonett, *Some Considerations of the Propagation of Electromagnetic Radiation through an Ionized Medium*, The Ramo-Wooldridge Corp. Mem. (9 Mar. 1955).
9. E. Langberg, K. Baldwin and J. Yos, *Radiation and Propagation of Telemetry Signals During Hypersonic Re-entry*, AVCO RAD, Wilmington, Mass.
10. Z. O. Bleviss, *Transmission of Electromagnetic Waves Through Ionized Air Surrounding Hypersonic Aircraft*, Douglas Aircraft Co. SM-22965, Santa Monica, Calif. (Oct. 1957).
11. I. W. Bond and J. N. Dyer, *Equilibrium Composition of Air at Various Densities and Temperatures*, Lockheed Missile Syst. Div. LMSD-1487 (Dec. 1955).
12. M. Van Dyke, *J. Aero. Sci.* **25**, (1958).
13. L. Lees and F. Kubota, *J. Aero. Sci.*, **24**, — (1957).
14. S. Feldman, *Hypersonic Gas Dynamic Charts for Equilibrium Air*, AVCO Res. (Jan. 1957).
15. E. C. Jordan, *Electromagnetic Waves and Radiating Systems* p. 681, Prentice Hall.
16. W. W. Balwanz, *Attenuation in Ionized Media*, private communications (3 Dec. 1958).

17. A. V. PHELPS, O. T. FUNDINGSLAND and S. C. BROWN, *Phys. Rev.* **84**, 559 (1951).
18. H. MARGENAU, *Phys. Rev.* **69**, — (1946).
19. P. MOLMUD, *The Langevin Equation and the A-C Conductivity of Non-Maxwellian Plasmas*, Space Techn. Lab., GM 47, p. 3 (16 Sep. 1958).
20. W. B. SISCO and J. M. FISKIN, *Effect of Relatively Strong Fields on the Propagation of EM Waves Through a Hypersonically Produced Plasma*, Douglas Aircraft Co., LB-25642, Long Beach, Calif.
21. A. S. DUNBAR, *Critical Pressure and Power for High-Frequency Breakdown*, Lockheed Missiles and Space Div., LMSD unpublished Memorandum, Sunnyvale, Calif. (12 Mar. 1959).

THE DILEMMA OF MAN IN SPACE

J. S. BRADY
Space Technology Laboratories, Inc.

Abstract—Economical system design requires that each component be utilized to its fullest capability with redundancy only where it is required. Man as a space vehicle component has unique capabilities for decision and repair which can be achieved by machine components only at great cost, if at all. Although design criteria for man-operated space equipment are stringent, techniques exist within the state-of-the-art capable of meeting the requirements of space systems.

INTRODUCTION

The design of man-operated space equipment has been the subject of a vast amount of highly fanciful discussion apparently arising from the unspoken premise that space represents an *unknown* which nullifies most, if not all, of our understanding of system design requirements. 'The Dilemma of Man in Space' is dedicated to the proposition that we *now* possess sufficient data and technology to achieve a workable, if preliminary, design based upon a selection between and extension of existing techniques.

A SYSTEMS ROLE FOR MAN IN SPACE

When the layman considers manned space vehicles he has a tendency to think, in the romantic terms of science fiction, of adventure, exploration and scientific discovery. The economist and sociologist view space travel in terms of cost vs. the returns from exploitation of resources and colonization; the strategist in terms of unlimited surveillance and invulnerability; the politician in terms of universal communication. The engineer, on the other hand, views the manned space vehicle as a problem in machine design. All of these viewpoints have at least one thing in common: they view man in space (MIS) as a payload, as something to be delivered to such and such a place in reasonable condition. To this end, a number of suggestions have been made to reduce the demands of the payload on the transporting system. Such suggestions range from the serious consideration of freezing or otherwise lowering the metabolic requirements of the passenger to the serio-comic suggestion that midgets be employed to reduce the size–weight requirements. These concepts of man in space have something else in common: they ignore a basic tenet of system design, the principle of *parsimony*.

An engineer, struggling with the problems of achieving escape velocity and astro-guidance, will recognize the fallacy of the payload concept when it is stated in the form:

"The space-weight requirements of any airborne (or space-borne) system demand that two criteria of component selection be met: (1) any component should be utilized to its optimum capacity and (2) duplication of components should be avoided except where there is a sufficient system reason".[1]

The man in space, as payload, is a component already specified for inclusion in the system. If the criteria are to be met, it follows that the automatization of any required function must be justified on the grounds of the man's inability to perform the function adequately. It is not sufficient, moreover, to demonstrate that automatization will result in superior performance. It is necessary to show, in addition, that the system requires this superior performance and that neither space-weight nor system reliability is unduly sacrificed for the increased performance.

From the practical, rather than theoretical, standpoint, since automatization of any function adds additional weight to the system and since the applicability of state-of-the-art guidance systems is at best limited, the time if not the fact

of successful space travel may very well depend upon the effectiveness with which the passenger is utilized as a functioning system component.

Man as a systems component has a number of well-known limitations. He is weak, variable and inaccurate compared to an electro-mechanical servo. He cannot handle as much data as a computer and is intolerably slow in calculation. He cannot be relied upon as a continuous monitor for appreciable periods. On the other hand, man has a number of unique capabilities which, although equally well-known, have been less explicitly enumerated.

Let us examine possible system functions and evaluate the relative efficacy of man and machine components. System related functions fall into three classes: intra-system, supra-system, and extra-system.

An *intra-system* component is one which functions as an intrinsic element of a system loop or chain, i.e. which contributes directly to the system output. Intra-system functions are implied by the terms "detection", "transduction", "computation", "actuation", "tracking", etc.

Supra-system components are those which, although they may determine system output, do not directly participate in the system function. Supra-system functions are characterized by "selection", "decision", etc.

Extra-system functions are those which are independent of the system transfer function, e.g. repair, fueling, etc.

It should be immediately apparent that the objections to man as a system component refer almost wholly to intra-system functions and, further, that machine capabilities in supra- and extra-system functions are extremely limited, and achieved only at great cost in time, space, weight and reliability.

A generalized planetary exploration, since it entails all of the essential elements of a space mission, will be used here as the model against which to establish appropriate en route functions for MIS.

Assuming a von-Braun-type torchship such a mission would involve:

(a) *Embarkation*, from a space station, consisting of preliminary, castoff and course orientation maneuvers, and of initial blast timing.

(b) *Astrogation* (probably by star track) and course–speed correction.

(c) *Planetfall*, including detection, surveillance, pilotage (site selection and acquisition) and landing.

(d) *Liftoff*, including time, course and speed establishment.

(e) *Stationfall*, entailing detection, acquisition and orbital maneuvering.

Paralleling these mission phases in time and the *reactive functions* of monitoring, emergency detection, emergency reaction and repair.

The mission and reactive functions can be seen to fall readily into a four-fold classification system. Lumping high speed, sensitivity, etc. under the term "precision" and freedom from established programs, purposive selection, etc., under "flexibility", we have

Precision		High	Low
Flexibility	High	Category *A*	Category *C*
	Low	Category *B*	Category *D*

Categories *B* and *C* can be seen to embrace the majority of important mission functions, e.g.:

Category B (high precision, low flexibility)
Course orientation and initial blast, astrogation, liftoff, detection, orbital maneuver, etc.

Category C (low precision, high flexibility)
Surveillance, selective acquisition, pilotage, repair, etc.

If we analyse the capability–cost features of our manned space vehicle, we find that they also can be fitted into the same classification scheme. The machine elements possess the precision, but lack the multiple programming and purposive selection of the human operator. MIS, on the other hand, exhibits the necessary flexibility but is deficient in the more mechanical functions. Allocating functions on this basis, Category *B* becomes the principle domain of the machine elements, Category *C* that of MIS.

Unfortunately, some functions, e.g. contingency aspects of landing, emergency reaction, etc., may fall into Category *A*.

Category *A* functions, those requiring both precision and flexibility, can only be implemented by attaining a critical integration of man and machine capabilities. They, therefore, pose the most difficult problems of system design. These functions are of two types; functions demanding breadth, sensitivity and/or duration of vigilance coupled with flexibility of reaction; and functions combining interpretation of situational implications with the need for high-speed, precise reaction. However, if the machine supplies the vigilance and *warning* in the first instance, reactive functions can be relegated to Category *C*. Similarly, if the man has the ability to *override* programmed performance, reaction in the latter situation becomes a Category-*B* capability.

Summary: Economical system design demands that functions involving situational interpretation, program override, purposive reprogramming and reaction to unprogrammed events be allocated to MIS.

operator's performance will be limited by the amount of task information available to him. Further, that information which does not contribute to that performance is gratuitous and may even lead to degradation by overloading his capability to process and select data.

Different forms of information are required for intra-system (control) and supra-system (decision) functions. These different requirements have resulted in the evolution of two major classes of dynamic displays which may be characterized as "end-point" and "situational" displays.

An end-point display exhibits a desired performance goal (end point) and the system deviations from that goal (error indication) or control action necessary to achieve the end-point (command indication). A situational display, in addition to showing end-point information, places the end-point and the system in some frame of reference, the action environment, thereby providing not only deviation from some pre-computed program, but a knowledge of environmental relations and events.

Fig. 1. End point and situational displays: (a) radar fire control display; (b) computing optical sight.

INFORMATION AND CONTROL SYSTEMS

Having arrived at a suitable systems role for MIS, it now becomes important to examine some relationships between current technology in equipment design and his ability to perform his assigned functions.

Design of Information Systems

It is perhaps superfluous to point out that an

The underlying distinction between end-point and situational displays will be easily visualized in comparison of the information presented by a contemporary aircraft radar fire control system and that displayed by a computing optical sight (see Fig. 1).

Although the fire control display presents the necessary information (course-to-steer, range, closing rate, etc.) in a highly precise and simplified precomputed form, the pilot is deprived of

information regarding the true action environment. Since he is given only a steering error, his role in the system is *restricted* to that of a servo-follower with its intrinsic lags and resultant tendency to oscillate. He does not have the information to perform in any other fashion. With the optical sight, on the other hand, although the pilot must rely upon size for range and rate of change of size for range rate, his computed information (reticule) is imbedded in the spatial and temporal environment of the system and its target. The pilot may avail himself of the relative, historical and instantaneous data present to anticipate target evasive maneuvers and/or to select alternative courses of action. It should be clear that the majority of functions allocated MIS require this latter type of display.

Amount and form of information displayed, although of great importance, do not comprise a complete specification for an information system. Display dynamics (control-display and performance-display relations) must be matched to the human operator if maximum utilization of his capabilities is to be attained. The necessity for performing complex interpretive and computational operations must be removed from the operator by taking advantage of his natural response propensities, by integration of information parameters and, where feasible, circumvention of control-performance lags supplying the operator with some degree of immediate knowledge of the eventual results of his control action. Since the goal of "unburdening" and "quickening", as these techniques have been termed, is to enable the operator to concentrate his attention upon the decision aspects of his task by relieving him from the routine aspects of performance, these considerations comprise an important part of system design for MIS.

Design of Control Systems

The effectiveness with which MIS can be expected to fulfill the functions allocated him will depend not only upon the information furnished by the system, but also upon the amount of control responsibility he is expected to assume. Control responsibility has two important implications for the manned space vehicle system. As MIS takes over more and more control functions, less automatization and, hence, less space–weight is required. On the other hand, increased control responsibility places a load upon MIS with a consequent reduction of decision capability. One of the important problems in the design of an effective control system is the establishment of an appropriate balance between these two factors.

If MIS is to serve as a *pure* decision operator, all that the system should require of him is that he select the appropriate program to follow. Complete responsibility for program execution would rest with the machine components. Needless to say, few practical tasks are sufficiently invarient to permit such push-button operation. Moreover, the greatest utility of MIS is achieved in situations for which no programs are available. Control systems appropriate to these situations require that the operator more or less directly implement his own decisions, thereby avoiding the need for a large number of pre-determined programs. Autopilots for aerodynamic vehicles supply control of this sort. The autopilot provides flight stabilization, aids in making maneuvers and maintains the flight conditions established by the pilot. The pilot remains primarily a decision operator with the added function of program generation and primary execution. Increase of control responsibility beyond this level can only be accomplished at the expense of operator decision functions, with little saving in space–weight. Since, even where the operator is required to perform more directly as a system controller, an irreducible amount of equipment is required for functions beyond his response capability, e.g. control boost, yaw damping, etc.

Regardless of the degree of control responsibility imposed, control system design must take into account his characteristics as an information transducer and decision generator. Current theory [2] views the operator as a probability computer who attributes increased likelihood to certain members of a pre-established set of alternatives on the grounds of information

received. To the extent that preprogramming is not available, the system must be designed to assist the operator in generating and implementing appropriate decisions. However, most tasks require at least a partially intra-system role for the operator. For this reason, it is desirable to examine in somewhat greater detail some of the characteristics of the operator which will have a direct influence upon the control system design.

Attempts to arrive at an accurate transfer function for the operator in this role have unearthed an almost unique fact concerning the operator: *he changes his transfer function to fit the demands of the situation.* This must not be taken to mean that optimization of the man–machine linkage can be to a large extent ignored. This flexibility is not accomplished without effort—effort that may produce overloading in critical situations. Optimal utilization may be achieved by taking advantage of the multi-function characteristic of the operator only when absolutely necessary, utilizing the simplest functions wherever possible.

The simplest transfer function, and the one that appears to be optimum for the operator, is that in which a force proportional to the magnitude of the error is applied. To consider the operator's output strictly as a force, however, represents a needless refinement—it may, in fact, be inaccurate in some cases. The operator's response to a step-function input, for example, shows a definite ballistic phase of approximately the same duration regardless of the amplitude of correction made. Thus, it may be desirable to approximate the operator output by the more convenient concept of a sequence of discrete displacements. To the degree that the operator makes corrections of a ballistic nature—i.e. sequential incremental responses—it would be expected that accuracy of performance will suffer when there is any appreciable delay between control action and indication of its result. This is the case with the human operator.

A further functional characteristic of importance which has been argued for the operator is that he is effectively a single-channel sequential computer, performing operations on a time-sharing basis. Protection against overload then requires that, when multiple tasks are to be performed (in effect, simultaneously), no single task requires continuous attention and/or correction.

These operational characteristics impose a set of minimum requirements upon the control system which must be achieved if maximum utilization is to be attained. These are threefold:

(1) The operations required of the operator must be limited to simple amplification or, at most, first-order integration.
(2) Time delay between response and result indication must be minimized.
(3) The responsibility for continuous correction must be removed from the operator.

A number of design techniques are available to meet these requirements.

Simplification of the control function for situations in which the desired end-condition is continuously changing may be achieved by a technique that has become known under the name of *aided tracking*. This arrangement may be considered as an attempt to combine the advantages of positional and higher order control. In idealized form, force applied to the control displaces the controlled member by a corresponding amount and also gives it an increment of the higher order derivatives. For the simple "rate-aid", it involves a feed-forward loop, and the linear relationship is given by

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \lambda\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right) + \mu\phi$$

where θ and ϕ are the displacement of the controlled member and the control, respectively. There are two ways of looking at an aided system. It may be regarded simultaneously as "*basically a velocity control*, but made effectively positional for small corrections, thus reducing its order and making it easier to operate"; or "*basically a positional control*, but with a velocity component added to avoid the necessity for continuous correction".

It should be noted that the effectiveness with which aided control performs within a given

system will depend upon the aiding constant, λ/μ, selected. This constant, in turn, has been found to differ within wide limits in different applications and consequently should be empirically determined.

Minimization of the control–display lags in those instances in which they result from processes inherent within the system may be achieved by a variation of the aiding technique. If we schematize a second-order control system, as in Fig. 2a, it will be seen that, so far as the

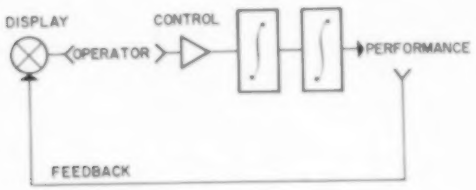


Fig. 2a. Second order control system.

display-control relation is concerned, aiding and partial circumvention of the time delays can be accomplished by feeding-back into the display the earlier-occurring intermediate functions, as shown in Fig. 2b. In the definitive

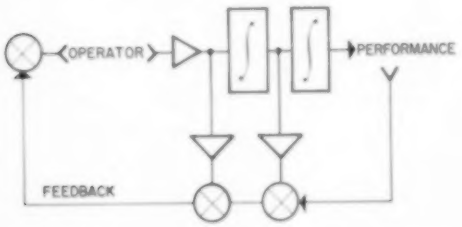


Fig. 2b. "Quickened" second order control system.

sense, any *aided* system is quickened; however, the term and the techniques are particularly applicable when the operation is performed upon the display–control loop rather than the control–output loop.

More direct and complete circumvention of the control–display lag may be achieved by use of an extrapolative device that displays, at any instant, the result of a given control action projected to some time in the future. Two forms of prediction devices with regard to control may be distinguished; those displaying a prediction over some fixed interval of time and, those displaying the steady state that will result if the instantaneous control action is maintained. All predictive devices, particularly those that perform their operations upon present performance plus command alone, suffer from the influence of uncontrolled parameters, to a degree proportional to the interval over which performance is extrapolated.

Additional control and display characteristics will be determined by the functional environment within which MIS must operate and must be predicated upon an understanding of the nature of this environment and its effects upon operator performance capability.

THE FUNCTIONAL ENVIRONMENT

The design of any system must take cognizance of the environment within which it must perform. In the present case this is the "in-space" environment of our human operator.

The ecological and psychiatric problems of this environment will be omitted from the discussion, not because they are unimportant but because their solution must precede consideration of control equipment design.

Insofar as the performance of MIS is concerned, the critical properties of the in-space environment are the extremes encountered within the acceleration and visual parameters.

MIS must operate in an acceleration environment for which his muscles and receptors were not designed, with consequent degradation of his "normal" response capabilities. The primary effects of "high-G" are the loss of capability to perform fine control actions, to make major limb movements and to sustain effort, coupled with blurring or loss of vision. Under "null-G", on the other hand, loss of normal gravitic cues (kinesthesis) produces a tendency to "overshoot" or "oscillate" in major limb movements and requires constant and immediate visual monitoring of results. There is some evidence that performance in the "null-G" case will improve with practice but how much improvement and how much practice are as yet unknown.

Three characteristics of the in-space visual

environment render precise control by direct vision extremely questionable. The speeds and distances involved require detection and acuity beyond the resolution of the unaided eye. Further, the brightness ratios encountered in the in-space environment exceed greatly those to which the human eye can adapt. As a consequence, unaided observation of an external "bright" object will be followed by a loss of acuity for internal (panel) objects until adaptation has taken place. Hence, normal sequential fixation (scanning) is not feasible. Further, the primary cues for depth perception and relative motion, i.e. those resulting from the presence of a continuous, textured ground, are absent from the space field. The remaining cues: size, rates of relative motion and size change, are inadequate for precise maneuvering.

A further functional problem, impulse control, is imposed more by the anticipated vehicle design than by the in-space environment itself. The difficulty lies in the fact that in-space maneuvering will probably be accomplished by ballistic propulsion (pointing and firing) rather than by continuous control. The control problem is essentially the same as that involved in the launching of unguided airborne rockets. It entails prediction of a collision or near collision course from relative locations and motions. As has been shown in fire control system design studies, precise performance in this task is extremely difficult if not beyond the capability of the human operator.

DESIGN CRITERIA FOR MAN-OPERATED SPACE EQUIPMENT

Our discussions of MIS roles, information and control systems, and of the functional environment have formed the background for a set of criteria upon which the design of man-operated space equipment can be based. The equipment with which we are concerned is chiefly that necessary to support those functions which we have defined as "supra-system" and "extra-system", i.e. supporting the functions of surveillance, pilotage, selective acquisition, warning and override control and repair.

1. Design Criteria for Surveillance

To combat the environmental restrictions, surveillance equipment should provide MIS with

(a) Extension of visual acuity.
(b) Situational information.
(c) Flattening of visual contrast.
(d) Enrichment of the visual field . . . perhaps by means of a fixed or synthetic/dynamic grid.

2. Design Criteria for Acquisition

Equipment for selective acquisition should supply

(a) Increased detection sensitivity coupled with warning.
(b) Situational information and/or positive end-point identification.
(c) Positive control/display acquisition relationship, e.g. an aided control-linked cursor.

3. Design Criteria for Warning

Above all, warning equipment must provide a sensory override, i.e. positive interference with ongoing sensory processes plus an identification of the nature and, where possible, source and locus of the aberrant function.

4. Design Criteria for Override Control

Override control equipment poses the most complex design problem. The fulfilment of this function requires that the equipment supply:

(a) Timely warning of the aberrant condition.
(b) Situational information including present performance, the consequences of continuing present performance and overall performance environment.
(c) Aided control–display including the necessary predictiveness to circumvent sensing lags and complex operator intra-system computations.
(d) Controls which do *not* require either movement of major body members or fine

sensory-motor manipulation, e.g. force stick or push button.

5. Design Criteria for Repair

In order that rapid and efficient repair of malfunctioning or damaged equipment may be accomplished, the space vehicle must be so designed and monitored that it supplies

(a) Continuing vigilance and immediate warning of malfunction and/or damage.
(b) Sufficient information to permit rapid diagnosis and localization of trouble.
(c) Access to critical components.
(d) Component size and design within the manipulatory capability of crew members under "null-G" conditions.
(e) Tools and handling equipment capable of being used in "extreme-G" environments.

Stringent as some of the design criteria appear to be, they present no problem which has not already been encountered and dealt with in the design of aerodynamic control systems. Ground laboratories exist, or are feasible, for the exhaustive investigation of all but "null-G" phenomena which can be generated for brief periods in aircraft. In short, nothing precludes the immediate initiation of design studies but the lack of directed effort.

SUMMARY

(1) Economical systems design requires that each component of a manned space vehicle be utilized to its fullest capability, with redundancy only where it is required.

(2) Man as a space vehicle component has unique capabilities for decision and repair which can be achieved by machine components only at great cost, if at all.

(3) Man-operated information and control systems must be tailored specifically to the functions and capabilities of the human operator.

(4) The "in-space" environment has a number of characteristics which impose tasks beyond the unaided performance capability of the human operator.

(5) Although design criteria for man-operated space equipment are stringent, techniques and designs exist within the "state-of-the-art" capable of meeting the requirements of space systems.

REFERENCES

1. J. S. Brady, *Aeronaut. Engng. Rev.* **14**, 86 (1955).
2. H. P. Birmingham and F. V. Taylor, *NRL Rep.* 4333 (1954).

A HYPOTHETICAL MISSION TO SPACE IN A THREE-MAN SEALED CABIN

J. I. R. BOWRING and B. P. EBERT
Box 7325, Area "B", Wright-Patterson AFB, Ohio

Abstract—This paper deals with a hypothetical flight into orbit in a sealed crew cabin with three persons constituting the crew. It is assumed that the vehicle is vertically launched and enters into orbit for 30 days followed by a high life re-entry and landing. Some of the problems concerning the accelerations during launch, weightless flight, cabin layout, high acceleration crew seats, restraint systems, feeding devices and human elimination devices are discussed. The paper endeavors to describe some of the WADC "in-house" efforts to find practical solutions to these problems.

This paper will discuss some of the factors which will face the crews of future high performance flight vehicles and give an idea of the work in this area being currently undertaken at the Wright Air Development Center. As many readers will be aware, basic research and development carries with it so much of the unknown, that often conjecture is mandatory; it is some of this conjecture that this paper will discuss.

For years, crew stations for aircraft have been a bone of contention. When the very early designers such as the Wright Brothers, Bleriot and others designed their flying machines, they first built the airplane, then decided where to seat the crew. It is now time to change this trend and design the crew station at least concurrently with or even slightly ahead of the flight vehicle. Engineers, designers and pilots have all attempted to come up with a good optimum crew station but to this day, we still argue about instrument location, color, arrangement, size, and so on.

To design a crew station, we must know something about its mission; and so for the purpose of this paper, we will consider an orbital vehicle with a crew of three, vertically launched with multistage rockets providing the power sources, in orbit for 30 days followed by a high lift re-entry, and a conventional landing on a gear or skids. To perform such a mission, the crew will be subjected to several phases as follows:

Phase (1). Pre-training and acclimatization.

Phase (2). Pre-flight checks, physicals, final check.

Phase (3). Launch vertically, the vehicle being subjected to accelerations of about 10 *G*, with vertical and longitudinal accelerations.

Phase (4). Orbit: weightless flight, mission accomplishment, feeding, elimination, entertainment, rest facilities.

Phase (5). High lift re-entry. Navigation and landing.

Fig. 1 shows the interior of a vehicle crew station. The cabin is 10 ft in length, and 78 in. in diameter, having a cylindrical section. Its capacity is a little greater than 300 ft^3. The crew is dressed in lightweight flight suits, not pressure suits, as will be the case for early flights. The cabin has no transparencies with the exception of the forward facing periscope which is used solely for the landing phase of the flight. This instrument will have a dual capability by presenting navigational information during the orbital part of the flight.

Though the crew can move about without the use of a floor, a normal environment will be desirable to meet the difficulty of orientation with fixed objects such as instruments, controls and other crew members. Floor, side walls and ceiling differ, however, from that of a conventional aircraft in that they are lined with a slow recovery rate material called "insulite". The complete interior is decorated with a balanced color scheme to provide orientation.

A sky color at the top, light brown on the walls and an earth color on the floor. The demarcation is pronounced to assist in establishing an orientation situation.

The crew of three are all trained to an equal standard, i.e. pilots, navigators, technicians and so on, but one is in command. The pilot is seated in the left seat, the co-pilot–navigator in the right, the off-duty man in the third seat or in bed. He will only occupy the third seat while attending to hygiene and feeding. At this time it is not determined what the optimum length of duty time or duty cycle should be, and it is hoped that work which will produce answers to this question will be undertaken in the near future. However, at the times for duty changes, which may occur at the end of every second orbit, the co-pilot goes off duty and moves to the third seat or bed, the pilot becomes co-pilot–navigator, and the man who has been off duty becomes pilot.

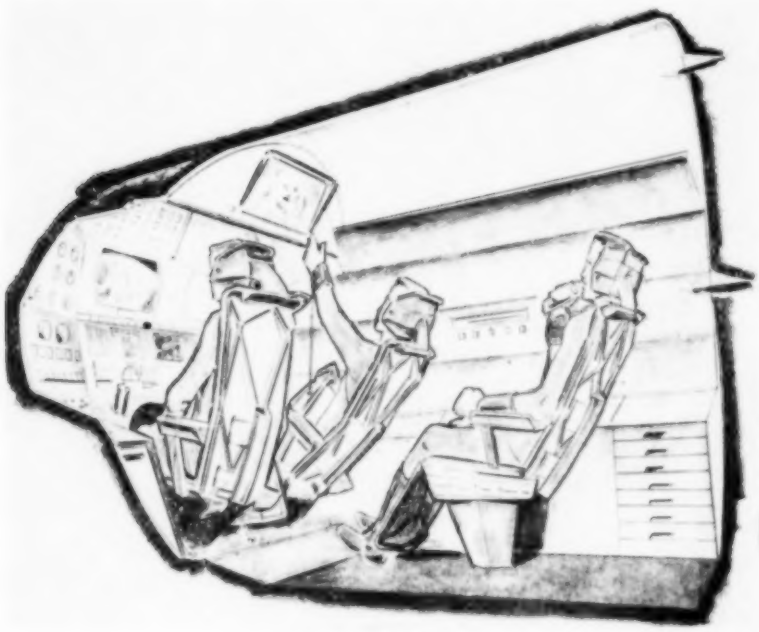


Fig. 1. Vehicle crew station.

Facing the vehicle operator is the instrument panel containing instruments grouped into those required for aerodynamic and space flight, systems management and electrical. A master warning/caution system occupies a predominant position and will give notification of impending abnormal situations. To provide a recreational facility, a small projection screen is available to show films, television shows and books which have been microfilmed. This device will also project technical instructions, etc.

How is the crew of such a vehicle to be fed? Highly condensed concentrated foods packed in toothpaste tube-like containers are good for short missions, but to most people it would be highly undesirable to eat unappetizing non-descript foods from such containers for periods of several weeks at a time. It is a most important factor to feed this crew correctly for the purpose of trying to create a normal environment. The average weight of daily food intake is 4·5 lb/person and requires storage space of 0·2 ft^3. Therefore, for a 30 day mission with a small reserve, say 10%, 435 lb of food occupying 18·2 ft^3 of space is required. How is the food

to be presented for consumption? It would be difficult to place food on a plate and eat with a knife and fork. The results would be catastrophic with food, such as French-fried potatoes and peas, distributed in many parts of the cabin when a plate was lifted and lowered—the contents of the plate having drifted off, unrestrained.

To solve this problem, it is likely that food will be packaged in thin rubber containers that must be capable of withstanding the temperatures of sterilizing food (250°C), must not give off a taste of rubber to the food and must be capable of being made airtight. Such a material

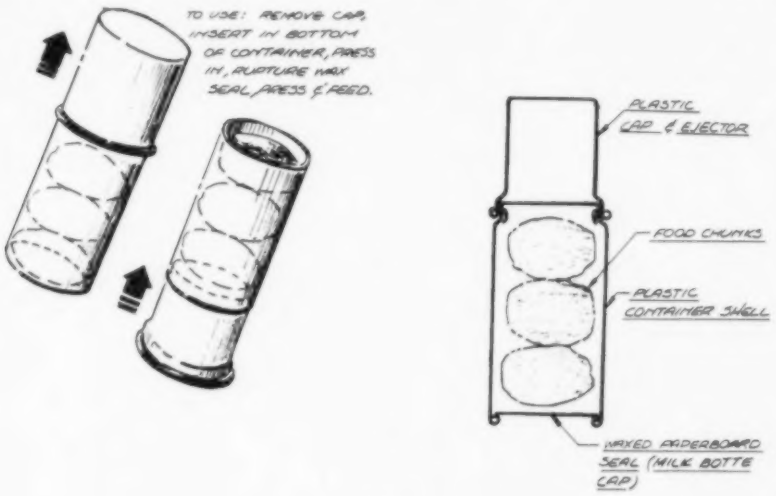
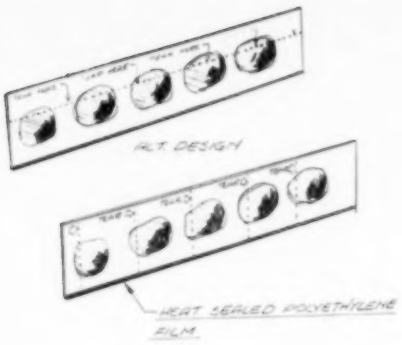

TO USE: REMOVE CAP, INSERT IN BOTTOM OF CONTAINER, PRESS IN, RUPTURE WAX SEAL, PRESS & FEED.
PLASTIC CAP & EJECTOR
FOOD CHUNKS
PLASTIC CONTAINER SHELL
WAXED PAPERBOARD SEAL (MILK BOTTLE CAP)

Fig. 2a. Zero "g" feeding device.

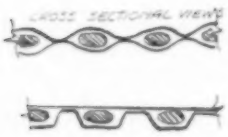
is available and known as nutrile rubber and is commonly seen as a seal in the caps of food jars. Using this material, we have experimented in manufacturing sausage-like containers capable of holding up to eight fluid ounces of food. These will be translucent and color coded. Such coding will indicate the contents of the tube as well as the consumer. To avoid mistakes and confusion, all consumable items will be coded to indicate their contents by meal type (breakfast, lunch or dinner), exact contents, and the container colored to refer it to a crew member. This system precludes the possibilities of conflict between the crew members resulting from dissatisfaction concerning food distribution. Storage of certain foods such as sandwiches or cupcakes will be by the use of individually wrapped items. This is similar to items seen on the commercial market today such as the packaging of crackers, cigarettes and other foods. The intent is to enable the individual to consume his food without physically touching it with his hands to minimize the dangers of infection, and also to prevent the man from discomforting contact with sticky foods. Examples of different packages are shown in Fig. 2(a–d). Of the total quantity of food involved, 30% is pre-cooked and then dehydrated; 15% is semi-solid, such as apple sauce and apricot sauce, then refrigerated; 10% liquids and refrigerated; and of the remaining 45% comprising meats and sandwiches, $\frac{1}{3}$ will be frozen. It will be seen, therefore, that provisions must be made for frozen (0°F $\pm$ 2°) chilled (32°F $\pm$ 2°) and nonrefrigerated foods. It is likely that the space between the inner and outer walls of the vehicle would form a convenient, if slightly unorthodox location for food storage. This storage area will be screened to guard against the hazard of radiation contamination to the stored food.

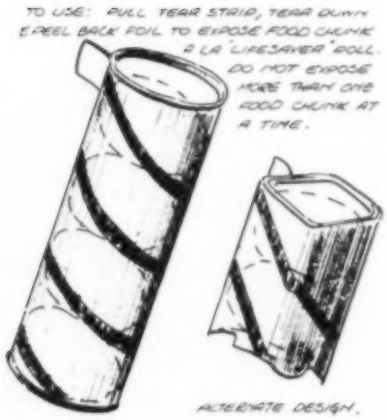
How to eat is the next problem. Quite certainly a conventional plate with a knife and fork will not suffice. The sausage-like containers which



TO USE: TEAR ALONG PERFORATIONS TO EXPOSE FOOD CHUNK IN INDIVIDUAL PACKET.

Fig. 2b. Zero "g" feeding device.




TO USE: PULL TEAR STRIP, TEAR DOWN & PEEL BACK FOIL TO EXPOSE FOOD CHUNK A LA "LIFESAVER" ROLL. DO NOT EXPOSE MORE THAN ONE FOOD CHUNK AT A TIME.
ALTERNATE DESIGN.

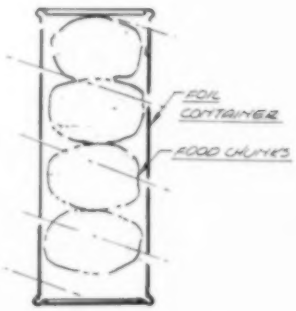
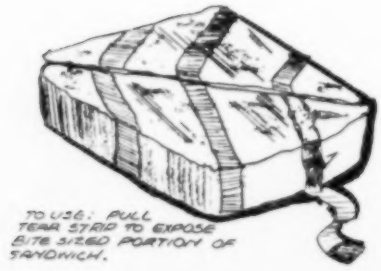


Fig. 2c. Zero "g" feeding device.



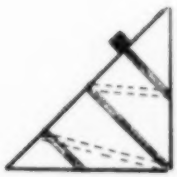


Fig. 2d. Zero "g" feeding device.

have already been described, are well suited for direct eating with the use of a suitable dispenser. We have developed two simple devices for this purpose. The first comprises a plastic container which houses a pair of small rollers spring loaded so that they impart a force between the rollers. These rollers are mounted in a slide which runs the length of the device. The rollers can be moved up or down this slide by operating a small endless belt. The outer container is hinged so that it opens in two sections, the rollers and slide on one side, and a cover on the other. The food container is now fitted into the dispenser so that the tongue at its end fits between the rollers and automatically engages on the engaging mechanism. The other end becomes a mouthpiece; being flexible it fits easily over the flapper valve and forms a sterile mouthpiece ready for eating by insertion into the mouth and pulling down on the endless belt. This device will not dispense any solid item of food unless it contains some liquid. However, this system is relatively complicated and needed some simplification so the next device, Fig. 3, was developed. A similar shell was used but instead of a pair of rollers to squeeze out the food, an air pressure system was used. One half of the casing holds a solid form with annular ridges along its length. The other half of the casing contains a flexible diaphragm with the same contour. When a food bag is fitted into the container and the diaphragm flexed, a pressure is exerted on the food bag which causes it to collapse extruding the food from the mouthpiece. The ridges on the diaphragm are so designed that the bottom of the bag, or end furthest from the mouthpiece collapses first. How do we collapse the diaphragm? A small sealed chamber, with a bellows and an inlet valve sensing cabin pressure, is provided for hand operation. The operator pumps the bellows, which increases the pressure within the chamber and exerts a force on the diaphragm, which as previously described causes an uneven pressure to be exerted on the food bag and subsequent extrusion of the food. The dispenser too will be color coded for identification. There are, of course, some undesirable characteristics when using any food bag type of feeder. A small study into eating habits reveals three main types of food selection by the consumer. These are:

(1) The person who consistently eats his food in a sequence—meat–vegetables–potatoes or some such combination. This type of person always completely eats one commodity before starting the next one. These people usually eat their preferred food last.

(2) The selective feeder. This person deliberates the sequence of eating and attempts to eat so that his food remaining is equal by type and quantity.

(3) The indiscriminate feeder. This man obviously feeds to maintain life and has little or no interest in what or how he eats.

The food bag system could eventually create a psychological problem for the Types 1 and 2, although if their individual habits were studied prior to flight, food could be packaged to accommodate their normal requirements.

Water is the next problem to be considered. With a diet of 2000 calories per day, a fluid intake of 5 lb of water is required at a cabin temperature of 68 $\pm$ 5°F and a relative humidity of 30 $\pm$ 5–10%. Some of this is taken in with the food, and some conserved as liquid. About 2000 cm^3 is taken by drink, the balance being used for reconstitution dehydrated foods. If we have to carry bulk water, 15 lb/day or a total of 495 lb including 10% reserve will be needed. Water reclamation devices are being developed and it is probable that a reasonably lightweight reclamation device will fairly soon be available for the recovery of water from urine, condensation and perspiration. Such a device must obviously weigh much less and in addition require less space than the water storage equipment that it replaces. How about body hygiene? Nature causes us to perspire and the odors given off are generally unacceptable to one's fellow man. How can we wash or keep clean—we can not use water for obvious reasons. For example, how could we contain the water so that we could wash with it and of course the problem of the additional weight of water for this purpose must be considered. Perfumed

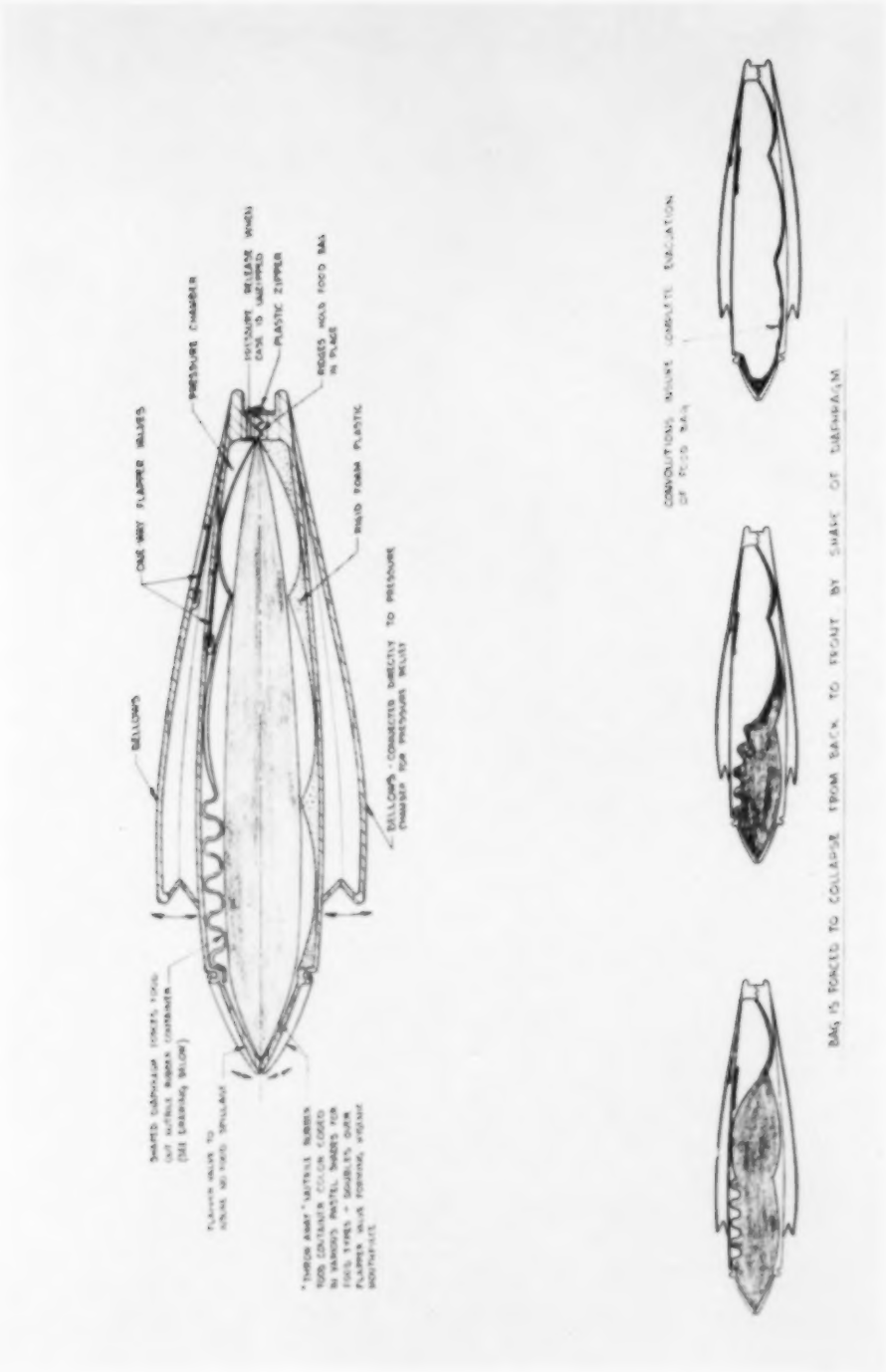

PRESSURE CHAMBER
ONE WAY FLAPPER VALVES
BELLOWS
RIGID FOAM PLASTIC
BAG IS FORCED TO COLLAPSE FROM BACK TO FRONT BY SHAPE OF DIAPHRAGM

Fig. 3. Zero "g" pressure feeder.

alcohol-impregnated wash cloths are the most likely substitute for water. This small pocket-size device not only cleans the surface of the skin, but refreshes it also. For those who desire to remain clean shaven, an electric shaver will be provided, with a vacuum attachment to collect the removed whiskers. It is customary as well as sanitary for the cultured human race to change clothing at least daily and this of course could be a problem. It was previously stated that the crew would be dressed in lightweight flying suits. Space and weight limitations make it quite impossible to take along 90 flight suits or one change per day per man. The likely solution to this problem will be to carry two single piece flight suits per crew member which have disposable underarm and crotch pads which will be changed as needed. A chemical deodorant will be used to treat the suits which are out of use.

It will be noted from Fig. 1 that the crew members are seated on a tubular frame seat. The escape concept for this flight vehicle is the recovery of the complete crew station so that an individual ejection seat or escape capsule is not required. A new concept in seating was needed, so a seat has been developed to meet these new requirements which are as follows:

(1) Position the body in its optimum position for launch and high acceleration. Maintain the body in this position at all times until conditions of weightless flight have been reached.

(2) Provide a total body restraint system such that the occupant will be restrained against acceleration, deceleration, vibration, transverse loads, gust loads and the forces encountered during a ground landing accident.

(3) Provide comfort for long periods of occupancy.

(4) Build adjustably to accept air force crew members from the 5th to 95th percentile of the service population.

(5) Have the capability of total encapsulation in the event of a major decompression of the crew compartment.

To establish the most comfortable position for the seated operator, it was decided to use nylon or Dacron Mesh, slung in moderate tension. A wooden frame with a series of $\frac{3}{8}$ in. holes drilled in it was used. The holes accommodate threaded rods $\frac{3}{8}$ in. diameter with a locknut at both the top and lower surface of the frame. The three main members comprise the back, seat pan and leg rest which can be adjusted angularly independently of each other. Nylon loops secure the nylon mesh to the test frame. Now by judicious manipulation of bolt length and angular adjustment, it can easily be seen that the seat becomes infinitely adjustable (Fig. 4).

The seat was rigged using existing known data on aircraft crew seats. Selected subjects were seated and after varying periods of time from 5 to 60 min, were asked: "Do you notice any discomfort?" When the answer was in the affirmative, the subjects were asked to denote the area of discomfort. The seat was then adjusted in an effort to improve its performance. All subjects were asked for comments on the general arrangement of the seat. After repeated adverse comment on a pressure point, the seat was adjusted until such time as the comment or pressure point was no longer noted. 264 subjects were processed through the experiment.

To establish both behavior and tolerances of the net seat system, a test program was arranged to subject the seat to acceleration, vibration and weightless flight. It was mounted in the gondola of the WADC centrifuge and a 200 lb dummy strapped into the position normally taken by the occupant. The centrifuge was operated so that a load of 3600 lb or 18 "G's" was exerted on the dummy, thus giving a proof run to establish the integrity of the prototype wooden seat. Subsequent inspection of the seat showed no distortion or damage and it was therefore considered safe to commence human testing.

An experienced and an inexperienced centrifuge rider were briefed for the initial run which was arranged so that the seat was semi supine, with the mean back angle to the horizontal of 20°. The first run was to be made with a rate of onset of 1 "G"/sec to 5 "G", with a sustained run of

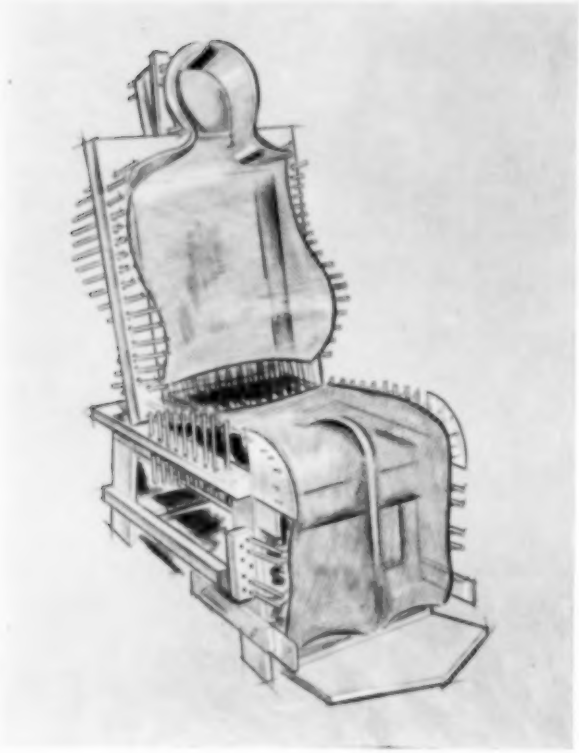


Fig. 4. "Breadboard" developmental "space" seat.

5 "G" for 10 sec and a run-down at 1 "G" for 5 sec. Both riders reported a comfortable ride, with the exception of some neck discomfort which appeared to have been caused by insufficient support of the neck, and remedial action which was in the form of a readjustment of the mesh. This involved increasing the tension locally and was made before further tests were conducted. Further runs to 5 "G" proved that the neck discomfort which had been caused by incorrect tensioning of the net solved the problem.

In view of the statement given by the riders concerning the comfort of the seat, it was decided to conduct some runs to high accelerations and using the same two riders, increase the peak accelerations to 9, and then 12 "G's" (see Fig. 5). At this peak riders reported the onset of chest pains.

It became necessary to determine the optimum angle of back to horizontal using the sling seat. It was unknown if the original angle used for

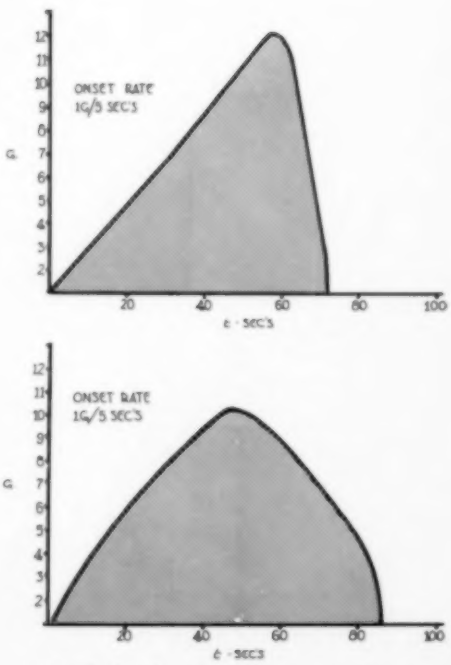


Fig. 5. Acceleration tests on two riders.

these experiments of 20° was optimum, and some further experiment was needed to determine the correct or optimum angle. The method for determination of this angle was to use the same riders for each series of tests and determine their reaction with regard to discomfort, chest pain and blackout threshold.

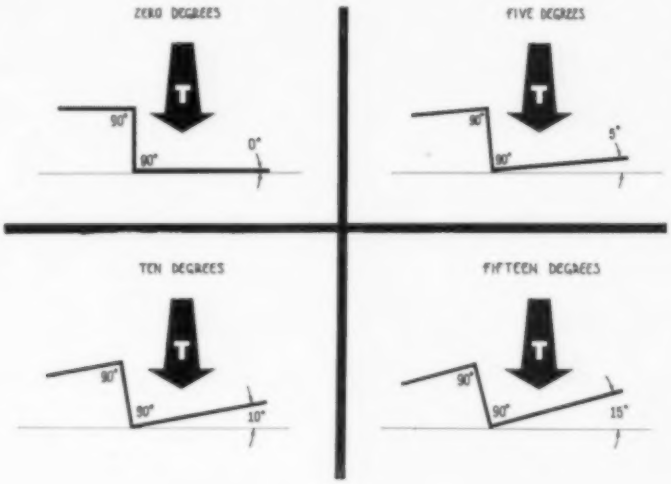


Fig. 6.

With the seat adjusted so that the mean back angle was 0°, 5°, 10° and 20° (Fig. 6), runs were made with observations concerning discomfort, chest pains and threshold. Clearly established by all riders was the fact that there was an optimum angle which produced a most noticeable lack of pain in the chest combined with a high "G" tolerance which exceeded previously accepted figures.

It was unknown how much stretch was being experienced with the nylon mesh, and so some runs were made to establish the true mean back angle at high "G's". A block of foamed plastic was mounted beneath the seat so that it was in line with the spine. Some 3 in. long nails were entered in the foamed plastic at 1 in. intervals so that about 2½ in. of the nails protruded. A ¼ in. square flat wooden block was glued to each nail. The contraption was arranged so that all nails were at a known distance from the STAFORM, and the subject's spine. The centrifuge was run to the high "G" figure and the body position was clearly shown by the now depressed nails. The mean trunk angle was then established for a range of subjects.

Whether Dacron or nylon rochel mesh to MIL-C-8061 would give the better results was determined by experiment. Samples of each were tested for elongation and tensile strength and the results shown in Table 1.

Table 1.

	Tensile lb/in.	Elongation per cent (to ultimate)
Nylon	120	118·3
Dacron	165	88·3

The seat was rigged with Dacron for a series of tests, and then rigged with nylon, followed by identical tests conducted using the same subjects. The opinions of all riders was that without question, the nylon was more comfortable than the Dacron. This is almost certainly due to the greater stretch of the nylon which provides more wrap around the chest walls during high acceleration. It ensures a greater contact area as the result of stretch which gives an increased capability of preventing possible rib distortion. All further testing was made using nylon mesh to MIL-C-8061

The next tests involved vibration. The resonant frequency, if any, was an unknown factor, and so it was decided to mount the prototype seat on a vibration table for low-frequency tests. Concurrently with this test program, work was in hand to develop a concept for body restraint for the occupants of the seat. Existing restraint systems with their restrictive harnesses and straps were considered too heavy—and uncomfortable. A restraint vest, made also of nylon mesh, was designed and a prototype fitted to the seat. The vest is essentially worn, the occupant sits in the seat and puts his arms through the arm holes, and engages the zip fastener. It can now be seen that he is wearing his restraint vest like a garment of clothing. So that the vest is completely adjustable, a battery of take-up reels are mounted behind the seat and are so arranged that the vest automatically adjusts to the size of the occupant. The take-up reels lock under positive load so that unless the reel lock is released, the occupant is firmly restrained in the seat. Fig. 8 shows the arrangement of the seat on the vibration table, so that the angle between the back and the horizontal is 20°. A series of thirteen tests were made with the same subject for all tests. At an amplitude of 0·5 in. and frequencies from 1 to 15 c/s in $\frac{1}{2}$ c/s increments the experiments were conducted and the first two showed serious problem areas. At between 5·6 and 6·8 c/s, the subject and the seat were out of phase with each other causing a bounce of the body.

With a 0·4 in. amplitude indicating on the seat, the buttock and stomach of the subject had an amplitude of 4 in., and this is quite clearly unacceptable. During the following weeks alterations were made which involved varying the net tension, adjusting the shape of the vest, but none of this improved the situation. The final solution was found to be in restraining the femur, which was subjected to a compressive load.

When this compressive load was placed on the knee, it put the femur in compression and effectively restrained the pelvis and stomach so that on subsequent tests at an amplitude of 4 in., the body had a maximum amplitude of 0·5 in. The test was considered satisfactory from the resonant standpoint. However, two interesting side effects were noticed. Many of the tests were conducted in the morning so that the subjects were not tired and more alert, but nevertheless, at the very low frequencies, 1–3 c/s, they were on several occasions lulled to sleep. At between 5·6 and 7·6 c/s the subjects were unable to speak intelligibly. It is assumed that the vibration coincided with a voice harmonic which caused this phenomena. This means that a voice transmission or radio would be impossible while being subjected to these vibrations. Further tests are planned in the near future, to vibrate the seat in both the horizontal and vertical plane concurrently, so that further research can be conducted.

It was decided that the prototype seat should be flown in the C-131, under conditions of weightlessness. The purpose of this experiment was to determine the relative position of the body to the seat frame when flown under weightless conditions and also the efficiency of the restraint system. The tests were uneventful with the exception of a noted fact that the seated operator was observed to move his legs in an almost horizontal position relative to the floor of the aircraft, when weightless. It is not known at this time how consistent this phenomena will be on a larger number of subjects, but it may well be that under conditions of weightlessness, this is a natural position for the legs. If this is the case, we shall certainly have to design equipment to permit man to adopt this position.

Where are man's arms to be placed when in the seated position? This is the next question to be considered. What are man's capabilities so far as an input to physically flying this vehicle using his arms, hands and legs is concerned? What controls should he have, where should they be placed? Certainly most of the controlling of this flight must be automatic but man's skill may be required for landing and perhaps on launch in the event that the automatic system fails or for other reasons doesn't cope with other emergency situations. In any event some type of up-down, left-right type of controller to take the place of a stick and rudder is needed in addition to on–off switches, variable controls

and so on. Tests on the centrifuge show that under conditions of high acceleration, precise operation and control of the digits, remain until black-out occurs. It is logical therefore to design primary controls for finger and thumb operation as far as practical. There is also a requirement to keep the arms off the chest during launch, because of the additional load imposed on the chest which would cause premature chest pains. An arm rest is a likely solution to this problem. A study of seated people in theaters, churches and at conferences revealed that most people sit with their arms either folded over their chest or with their arms crossed.

An arm rest which will permit the arms to sit in their natural position, over the chest was designed to permit easy entry and exit from the seat. The arm rests open from the center and hinge upwards and outwards. They have, also, a quick disconnect for emergency purposes such as a ground landing accident. On the arm rest will be mounted the primary control which will operate directly or indirectly the vehicle controls —pitch, yaw and roll is initiated or controlled by this controller. It is not mounted in line with the vehicle axis and is somewhat offset but preliminary experiments show that only a small amount of training is necessary to become proficient with this system. Of course the control will be statically and dynamically balanced.

One other problem with this control is that it is operated by the right hand only. In the event of incapacitation of the right hand, operation by the left hand is almost impossible, but as a multi crew operates the vehicle, in such an event, another crew member will take control. The modern trend of side sticks (X-15 and others) also brings with them this problem of what to do in the event of incapacity of the operator's hand and probably must be regarded as an accepted hazard. Other controls on the arm rest consist of a batch of variable finger controls operated by the left hand. These fit the natural finger tip position, as previously stated, and under acceleration loads, the fingers are capable of accurate positioning and are capable of operating this type of control. The arm rests also house an arm restraint and exercise device. During launch it is necessary to completely restrain the arms, and under condition of weightless flight it is most desirable to simulate a 1 "G" feel. It will be necessary to maintain muscle tone of the body so the system can be used to give up to a 3 "G" pull. Take-up reels mounted in the arm rests and connected to the wrists by cables are designed to lock the arms into the arm rest for launch. Depending upon the wishes of the occupant, this can be set from a loading of zero to 3 "G". A similar device connecting the ankle to a take-up reel is also used to take care of leg muscles.

By this time, most of the testing was complete, and engineering on the final configuration of the seat began. Using the contours and angles found by experiment, it was decided to manufacture a tubular titanium frame. When engineered for a 32 "G" capability, this frame weighs 7½ lb. Due to the high cost of titanium, and the difficulty of working it in our experimental fabrication section, we have constructed four frames using stainless steel and the weight of these is 10½ lb. The nylon mesh is aqua in color and is attached to the seat frame by means of a series of clips shown on the side of the seat. These clips pick up each second hole in the mesh; so it is possible very accurately to adjust the net as well as insure that the desired tension is readily obtained. After adjustment of the mesh, the raw edges of the mesh and the clips themselves are concealed, by the use of polyurothene protector strips. These also protect the crew members from the hardness of the metal frame in the event of minor collision, by a crew member with the seat.

The head rest is made of fiber glass with a foamed rubber insert. By providing several sized inserts, a close head fit can be readily obtained. Although in this model the insert is glued in place, the production head rest would have fasteners to secure the insert which would be a personal issue to the space pilot. A later model of this head rest will have to incorporate earphones and a boom type microphone.

The cords which provide tie downs from the mesh to the seat frame about the pivot point, are intromeshed. This allows for adjustment of

the seat fan relative to the seat back, or vice versa, so that the seat can be easily adjusted to accommodate the smallest or largest occupant. An adjustment mechanism will provide either manual or electrical means of adjustment. A lever operated manually moves the seat angle from the launch to normal work and rest position.

constantly wearing a pressure garment. There are so many practical reasons for not wearing one that we must consider an alternative system which will offer protection to the crewman at all times.

Encapsulation of each seat individually would appear to be a possible solution, and after

Fig. 7. Encapsulating system for lightweight crew seat. Normal position.

As previously discussed, there are many reasons for not having the crew wearing pressure suits for a flight of this duration. It can easily be seen that a serious loss of cabin pressure could well occur at any time, and might occur at a time when a crew member is partly undressed. What then if he is wearing a pressure suit? If the loss of pressure is rapid and complete, it is unlikely that the crewman will adjust his dress in time to prevent an explosive type of decompression from which he will be able to recover. Think also of sleeping, body hygiene, etc., whilst

encapsulation, an emergency supply of air will be used to pressurize the capsule to half an atmosphere with the occupant pressure breathing dioxyde through a mask. Such a situation, of course, in itself would present grounds for an abort of the sortie and immediate recovery of the vehicle. With the use of the most modern flexible nylon products, we are attempting to perfect a lightweight encapsulation system. The seat is surrounded by lightweight material which is stowed (Fig. 7), and when an emergency situation presents itself, the crew member must

occupy the seat, close the arm rest, and the seat will automatically be encapsulated (Fig. 8). In an emergency situation the seat is self-sufficient with its controls, pressurization and transparent panels which enable the seat occupant to look out of his capsule and see the instruments, periscope, etc.

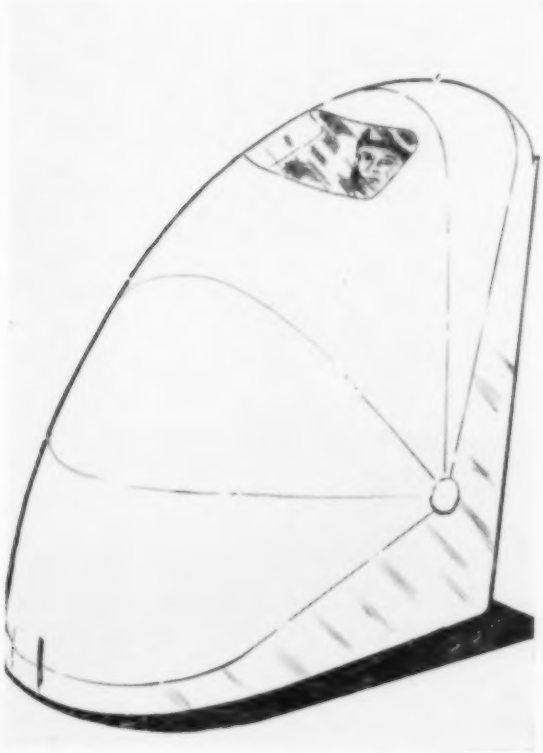


Fig. 8. Encapsulating system for lightweight crew seat. Enclosed position.

Let us return to the space cabin where our crew members are now in orbit and one of them is ready for sleep. During the prelaunch training period, the crew was taught a new living cycle which included an adjustment from the normal diurnal cycle. Consider for a moment, what would happen if we launched a vehicle this morning at 0900 hr, with a crew who had just spent the previous night at home. By normal retiring time tonight they would all be ready for sleep. The training period will take care of this, having readjusted the crew to a new cycle so that ideally one man is ready for sleep as soon as the vehicle is launched while the other two are out of phase with each other, one having only a half of a duty period to serve before rest.

In a state of weightlessness can we assist the crewman to orientate himself correctly by providing him with a normal environment as previously explained: a ceiling, floor, color definition, and now all that we have given the crewman in the way of assistance is gone as soon as he closes his eyes. Of course it is not certain that he will experience disorientation, but it is most likely that his vestibular senses will give false indications, and once again cause disorientation.

To discover if we could design a bed, with a device which would assist the occupant in knowing which was top and bottom, we conducted a simple experiment. A vertical wooden frame with shock cords mounted in tension was prepared. A net sheet was stretched over the top

of the shock cords so that a human could be sandwiched between the two. The subject for experiment was first of all blindfolded and then taken into a room in which the test apparatus was situated. He was put in the standing position, and the device tightened about him so that the shock cards were at the back and the mesh about his chest and abdomen. This experiment lasting 5 min each per person involved about 100 people. After this shock period of confinement in the frame, the subject was asked: "Do you feel more pressure at the front or back?" 96% could sense the greater pressure on the shock cord side, which for the purpose of the experiment, as previously stated, was to the subjects' back. Using this information we designed a bed which is 24 in. wide, and 73 in. in length.

To provide the minimum contact on the back, a stiff rubber corrugated mattress is provided which allows only the crests of the mattress to contact the body. The corrugations are upholstered with coarse mesh. The covering sheet which is made of elastic nylon is attached to four take-up reels. The occupant now has a pressure differential to recognize as well as the difference in the texture materials. It is expected that occupants will spend a pleasant period of rest on this bed. The bed has been designed to accommodate the off-duty crew members during periods of acceleration. Unfortunately the bed will not provide the optimum degree of "G" protection, but will insure a safe ride for the occupant. To permit the occupant to take acceleration from chest to back under all types of acceleration, the longitudinal axis of the bed has been installed perpendicular to the direction of normal acceleration. A simple "G" sensitive servo mechanism then rotates the bed to a position which will allow the acceleration to be taken from chest to back.

To provide relaxation facilities for the crew, we shall have a small 10 × 10 in. screen for projection of films, filmed television shows, micro-filmed books and technical orders. The forms of recreation will be dictated by the tastes of the individual crew members, type of mission and space available. Music will also be available on an "as required basis". Such items as the Chinese string game, checkers and chess will also be provided for use by the off duty members of the crew, although there will probably be very little time for participation in such matters.

How to take care of body elimination is the next problem. Urination will probably present little difficulty as urine must be collected for recovery and purified so that it can be fed back to the drinking supply. Therefore, urination must occur before and separately from elimination, so that no urine is lost in the elimination device. Elimination has to be assisted mechanically. It is most unlikely that the feces will have sufficient momentum to leave the body without mechanical assistance. Several devices have been proposed and I shall outline one such device. A seat over which a personalized cover is fitted before use, is mounted on a cylinder. A strap restrains the user on the seat. Inside the cylinder is fitted a double venturi, and air from the cabin circulation system is fed into the eye of the venturi causing a local flow and a down draft at the seat. Into the eye of the venturi is fitted a polyethylene bag. This bag has four small pouches which fit over four lugs, and a double chain string at its circumference. The lower end has sixteen $\frac{1}{32}$ in. holes to permit air flow. The user seats himself, secures the restraint strap, turns on the air supply and eliminates, the feces being drawn into the bag. On completion, the user pulls the draw strings which close the top of the bag and releases the strings. The bag and contents now are drawn by the air flow into a storage container (which is vented to the inlet side of the air conditioning system) through an odor valve and will remain in storage until the termination of the flight. After use, the occupant shuts off the air valve.

At the cessation of the mission, the vehicle reentered the atmosphere by using either high lift or high drag. The deceleration forces in the case of high lift are not excessive, about 1·6 "G" and such loads are comfortably taken in a back to chest direction. Deceleration forces in the case of high drag are much higher, in the order of 9·6 "G", and can be tolerated for a short period of time only. Nevertheless, these also will probably be taken back to chest.

After recovery, it will be necessary to retrain the man to live his conventional life under conditions of 1 "G", and the normal conditions on the Earth. This may take a week or so before he is properly readopted to his "natural" habitat.

This paper attempts to discuss briefly some of the early problems which have been considered, and to outline some possible solution to these problems. As the state of the art progresses, so ideas will change. It is paramount that those of us concerned with the development of crew stations, keep abreast of these problems and strive for improved solutions.

SPACE SIMULATORS—PRELUDE TO MANNED SPACE FLIGHT

A. B. THOMPSON and W. B. LUTON
Vought Astronautics, Division of Chance Vought Corporation

Abstract—Before man travels in space as an active member of a man–machine system, considerable additional knowledge must be gained as to his capabilities under the combined stresses of space flight and in the particular vehicle configurations being considered. Combination flight control–navigation–environment simulators are proposed as a tool to verify and aid analytical studies of man's performance. A review is made of two recent simulator programs (a centrifuge simulator and a six degree of freedom fixed base simulator) which studied human performance in orbital vehicles. Results of these studies are presented as examples of information on man's capabilities that may be obtained through simulation. Need for more complete environmental simulation is shown and a study presented to show the degree of simulation which can be reasonably obtained in such a specialized laboratory. One concept for a space simulation laboratory is detailed. Conclusions are drawn as to the urgency of need for such tools.

Before man ventures into space as an active member of a man–machine system, a thorough knowledge must be available as to his capability adequately to control and navigate the vehicle throughout the entire mission profile and to cope successfully with emergency situations. Although a considerable amount of data is available on human capability and tolerances when exposed to individual stresses, almost none exists for the multiple stress situations which will be encountered in space flight. Degradation of pilot performance under the combined stresses of noise, vibration, acceleration, high temperature, reduced pressure, isolation and extreme tension —to mention only a few—must be thoroughly evaluated and then compensated for insofar as practicable by optimum integration of the man with the machine. Theoretical analyses based on past tests establishing human tolerances to individual stresses are not sufficient to provide assurance that the pilot of an orbital vehicle will be capable of reliable decision, command and control throughout all phases of flight. The first rocket boosted flights of a manned space vehicle will be roughly equivalent to taking a new, very high performance aircraft to the extremes of its flight envelope on its first flight completely on instruments and without the benefit of a chase plane. It becomes increasingly apparent that relatively complex flight control–navigation–environment simulators must be utilized extensively from the study and design phases of vehicle development up to and including complete mission flights.

In the past, simulators have been used primarily for training purposes, being available to user activities in an after-the-fact fashion. Currently, they are being utilized on an ever-increasing scale for research in man–machine integration and optimization of vehicle parameters. However, the approach is in bits and pieces, i.e. pressure chamber type space cabins, heat chambers, acceleration via centrifuges and high-speed sleds, and fixed base flight simulators. Of course, these are logical first steps toward realization of manned space flight, but they fall short of determining man's usefulness under combined stresses and in the particular vehicles under consideration.

During 1959 two such simulators were designed, built and operated by Chance Vought Aircraft, Inc., as aids in establishing the needs and capabilities of the pilot of an orbital hypersonic vehicle. The first of these was an acceleration flight control simulator (Fig. 1) which operated both fixed base (static) and under boost acceleration fields in the Aviation Medical Acceleration Laboratory centrifuge at Johnsville, Pennsylvania. Its purpose was to determine the pilot's accuracy of control during acceleration representative of an orbital vehicle's exit flight, evaluate a new concept in pilot operated control

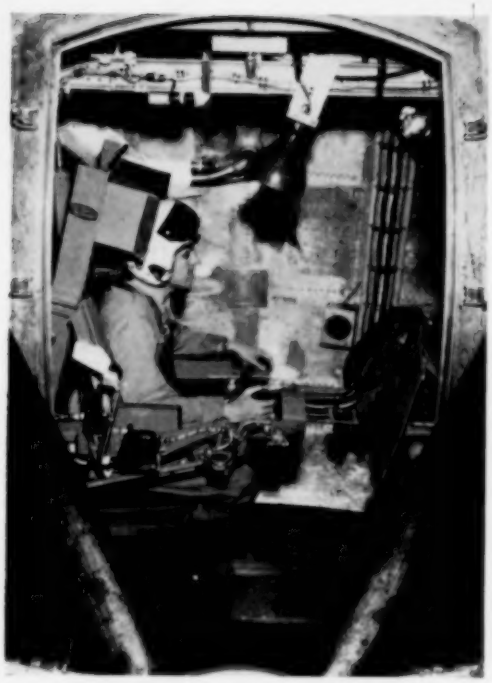


Fig. 1. Acceleration space flight simulator in AMAL, Johnsville, centrifuge.

seat position, and as rudder control became ineffective at approximately 6 "G" while wrist and finger control was effective to beyond 12 "G", a three-axis wrist controller would probably be necessary for manual vehicle control. However, all side sticks developed to this time were two-axis types—pitch and roll only. After studies were made as to wrist bone structure, pivot angle limits and force capabilities along all axes, a three-axis side stick was designed and initially evaluated versus the conventional center stick–rudder pedal arrangement in a C-11B simulator. With a minimum familiarization period pilots were able to make equally accurate instrument approaches and landings with either type of control. Both the selected seat angle and side stick were integrated into the centrifuge simulator along with the appropriate instruments necessary for the pilot to fly a command boost profile. The tracking task was a closed loop simulation of vehicle dynamics and response through the controls and instruments. Transient moments of wind shear and uneven thrust termination were introduced in pitch at

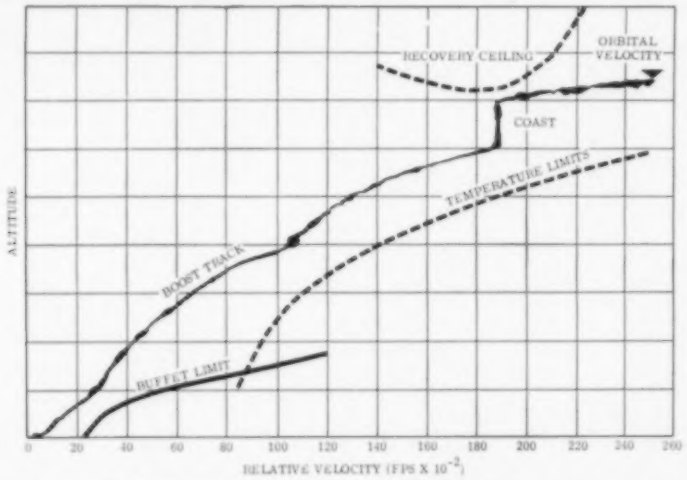


Fig. 2. Typical pilot control of vehicle boost trajectory.

devices and verify general physiological tolerance aspects to the rocket-boosted flight of a space vehicle.

Previous tests had indicated that optimum tolerance to high transverse acceleration fields was obtained with a slightly forward leaning

appropriate times for added realism. In addition to manually flying the entire trajectory into orbit the pilot was required to initiate and terminate last stage thrust.

Four experienced pilots flew twelve recorded runs each under 1 "G" conditions and twelve

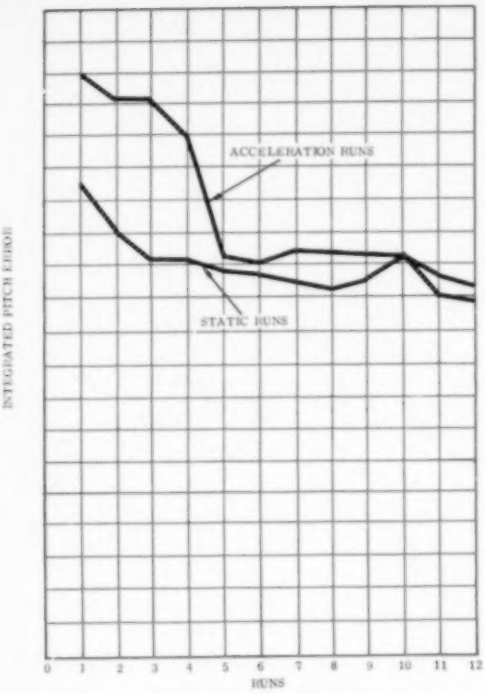


Fig. 3. Pilot performance degradation due to transverse acceleration.

flight path angle. Also a variplotter operated each run and plotted altitude versus velocity, a typical dynamic run being shown in Fig. 2. The four subjects' integrated pitch control error (from the commanded profile) was averaged for each run, static and dynamic, and a comparative plot made (Fig. 3). It indicated that after a short centrifuge learning period (four runs), the subjects' accuracy of control under the high "G" conditions was within 5 per cent of the accuracies obtained under 1 "G" conditions. The subjects' time of response to attitude errors averaged one-third greater under the boost conditions, however, this apparently did not significantly affect the probability of achieving a successful orbit. It was concluded that with the seat position and side controller used in this experiment and with proper flight information, a pilot could be an effective redundant controller to an automatic system during the boost phases of space vehicle flight.

As much of the remainder of an orbital mission was well beyond the regions which have been explored by and are familiar to engineering and

Fig. 4. Six degree of freedom fixed base space flight simulator.

recorded runs under the selected acceleration profile. Oscillograph records were made of integrated error in pitch, roll and yaw; pilot input in pitch, roll and yaw; instantaneous error in pitch, roll and yaw; radial acceleration; and

flight personnel as regards the human component in the system, a six degree of freedom fixed base space flight simulator (Fig. 4) was constructed. Specific objectives were to determine what control capability the pilot will have as well as

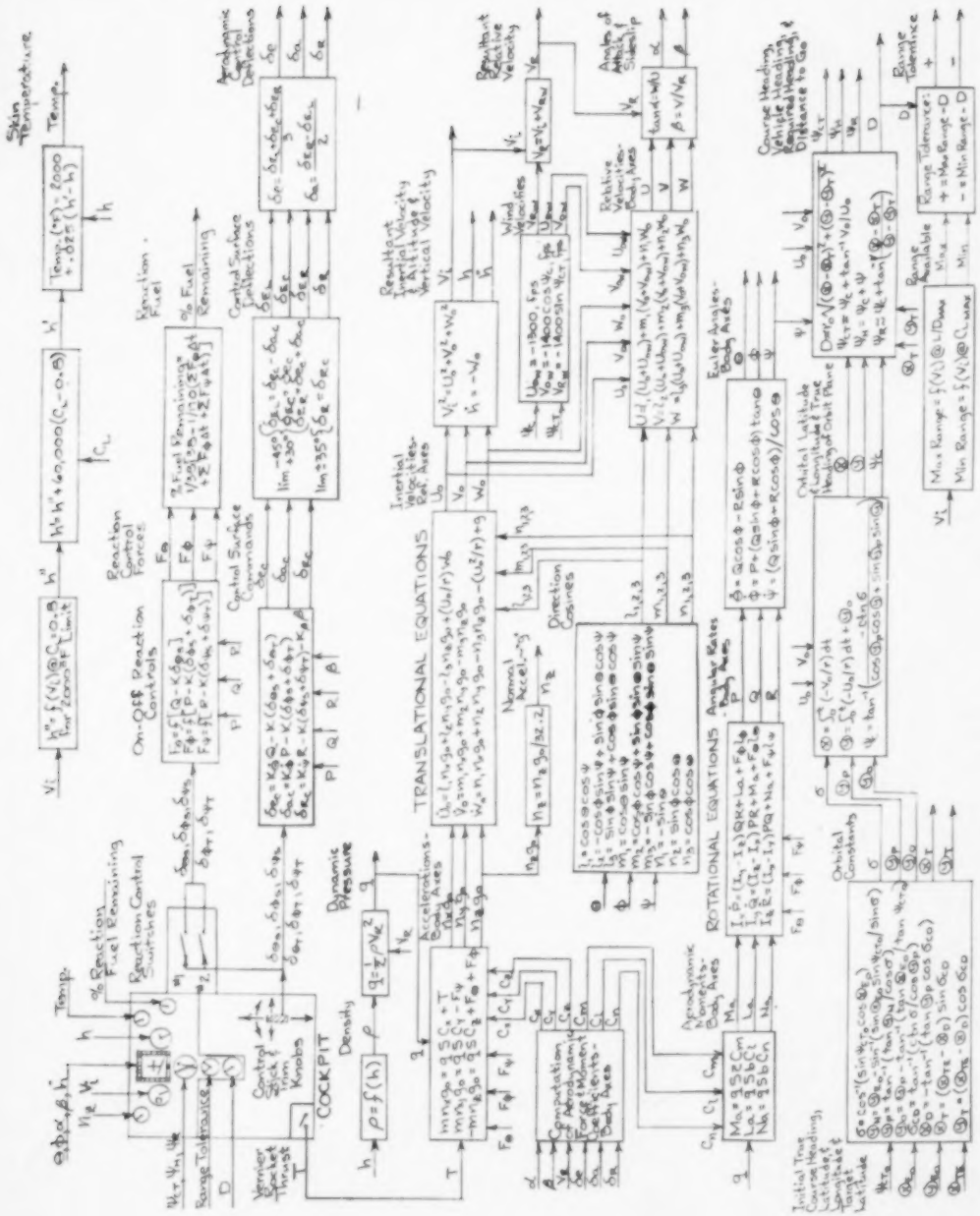


Fig. 5. Six degree of freedom space flight simulator block diagram.

what displays are required—both optimum and minimum—to manage energy, effect safe re-entry and navigate to destination. Some idea of the complexity required to achieve such simulation is illustrated in the block diagram (Fig. 5). This particular simulator was driven by 290 amplifiers and 16 function generators and was capable of simulating in real time the mission from the end of boost, throughout orbit, re-entry, hypersonic glide and finally supersonic approach pilot controlling temperature during re-entry and managing the vehicle's kinetic energy to arrive at destination. An orbital space vehicle requires a completely different concept of managing energy than pilots have used in the past. For example, at 500,000 ft altitude the potential energy (which is familiar to pilots) is 500,000 ft-lb/lb of vehicle weight, while the kinetic energy is twenty times that value or 10,000,000 ft-lb/lb of vehicle weight. The major

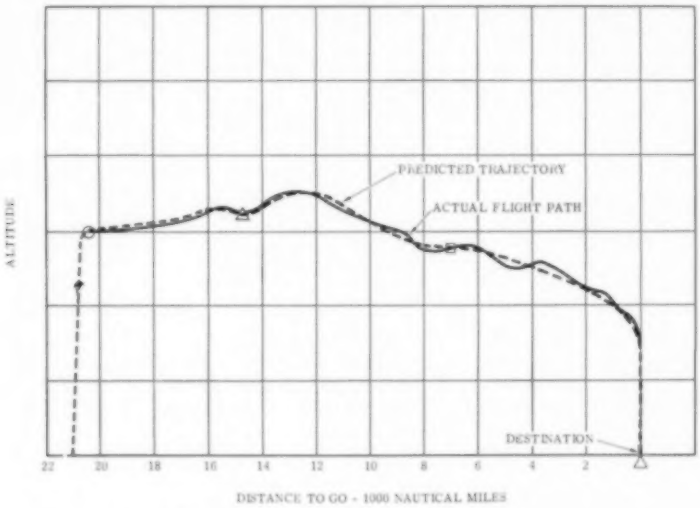


Fig. 6. Pilot ability to manually fly orbital mission.

to a point over the destination. All instruments were activated that were required for these phases—vertical situation, horizontal situation, distance to go, range tolerance, inertial velocity, inertial altitude, vehicle temperature, normal acceleration, reaction control fuel and time. The simulator contained the vehicle equations of motion, aerodynamic and reaction control parameters, long and short period dynamics, temperature and navigation equations, as well as full pilot control capability through the three-axis side stick.

In addition to standard pen recorder records of specific parameters, four variplotters recorded flight profile data. An altitude versus distance trace (Fig. 6) recorded the entire orbital mission as flown by the pilot. The deviations from the predicted trajectory are the result of the problem now is how to manage the velocity rather than the altitude. An energy management plot (Fig. 7) recorded inertial velocity versus distance. The pilot of an orbital vehicle may vary his range upon re-entry by as much as 6000 miles by varying the vehicle angle-of-attack between the limits of $L/D_{\max}$ and $C_{L_{\max}}$. Provided with an adequate display of this information, a wide range of destinations may be reached. Another parameter of critical concern during re-entry is vehicle temperature. A lifting vehicle (during a non-equilibrium re-entry) re-enters with a series of long period bounces off the dense lower atmosphere. A recording of altitude versus velocity (Fig. 8) illustrates this condition. As maximum lower surface temperature occurs at the bottom of each plunge and it is a function of angle-of-attack, it was found

that the pilot could effect safe re-entry from above the normal or equilibrium recovery ceiling. This was accomplished by holding $C_{L_{max}}$ during the descending phase and allowing the angle-of-attack to decrease as the temperature limit was approached and then holding L/D_{max} during the ascending portion of the bounce. Such a maneuver is relatively easy for a pilot and relatively complex and less reliable for automatic equipment. The ability of the pilot to turn off the orbital path during re-entry to reach particular destinations is illustrated in Fig. 9. With proper energy management and navigation information, the pilot may correct for relatively large azimuth launch errors or turn to reach alternate destinations.

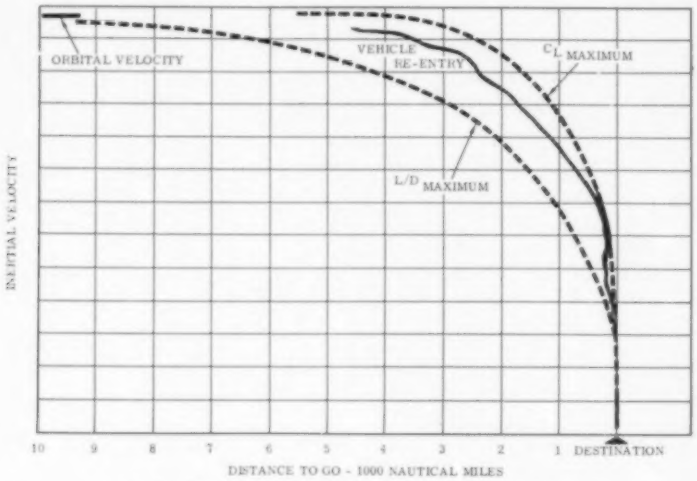

ORBITAL VELOCITY
VEHICLE RE-ENTRY
C_L MAXIMUM
L/D MAXIMUM
INERTIAL VELOCITY
10 9 8 7 6 5 4 3 2 1 DESTINATION
DISTANCE TO GO - 1000 NAUTICAL MILES

Fig. 7. Pilot management of re-entry kinetic energy.

Conclusions drawn from the operation of this simulator were that the human component, equipped with a display comparable to that used in the simulator, can manually fly the orbital, re-entry and hypersonic glide phases of a one-orbit mission and can manage kinetic and potential energy and navigate to various destinations. The capability of the pilot, assuming

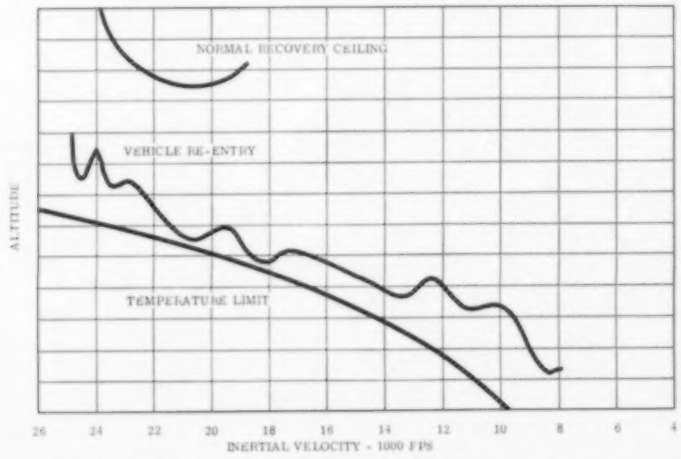

NORMAL RECOVERY CEILING
VEHICLE RE-ENTRY
TEMPERATURE LIMIT
ALTITUDE
26 24 22 20 18 16 14 12 10 8 6 4
INERTIAL VELOCITY - 1000 FPS

Fig. 8. Typical pilot control of vehicle temperature during re-entry.

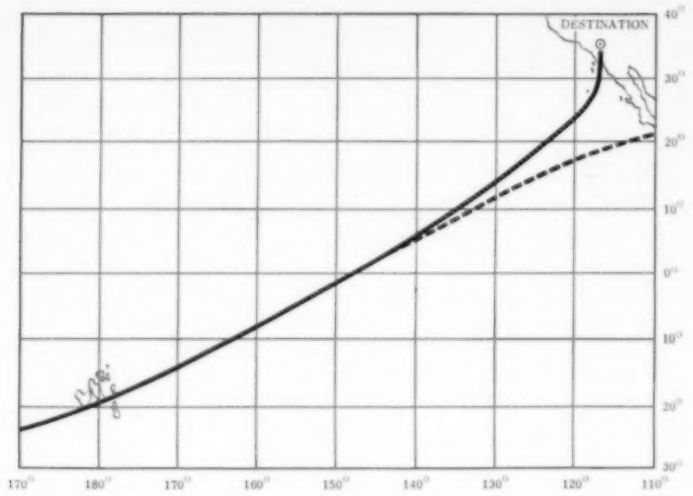


Fig. 9. Typical pilot controlled heading change to reach destination.

adequate training, is limited only by the performance capability of the vehicle. The only reservation in reaching these conclusions is that several other human factors—combined stresses, etc.—were not simulated and were therefore beyond the scope of the analysis.

simulation facility for every type of spacecraft is not necessary. Many of the parameters are the same for all crafts and missions, and many others differ so slightly that the apparatus would be readily convertible to each configuration. Studies were made to determine what degree of

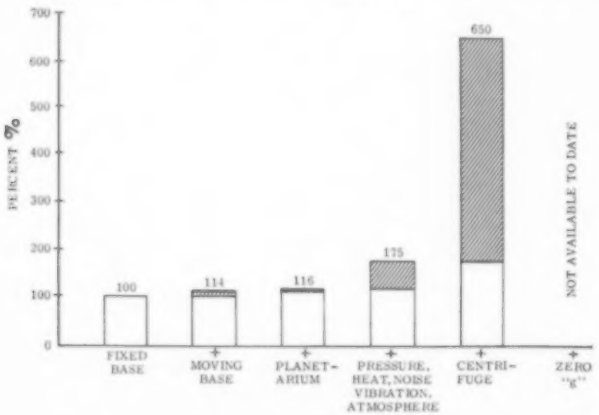


Fig. 10. Relative cost of degree of simulation for space vehicles.

Although the work to date on space flight simulators is significant as indicated above, it represents only a millesimal fraction of what is required before space flight becomes even relatively commonplace. As the next step, a most urgent need exists for a moving base environmental flight simulator. To build a near-total

simulation could be achieved at reasonable cost; the results of which are shown in Fig. 10. The basic fixed base simulator would simulate on a real time basis in six degrees of freedom all phases of the mission boost, orbit, re-entry, glide, approach and landing. The addition of a moving base would give the proper direction of

the acceleration forces existing at all times as well as the translational motions due to vehicle short period dynamics. External vision in the form of a planetarium type display would add very little to the total cost. The important crew environmental stresses—noise, vibration, temperature, internal atmosphere and pressure, and lighting—could be included for a reasonable price. Only when an attempt is made to include the magnitude of acceleration does the cost go out of reason. This is due to the extremely large centrifuge required to carry the heavy simulator. And, of course, the weightless condition is not obtainable at this time in surface-based simulators.

Conceptual studies were made to determine how best to achieve the above. One such concept is illustrated in Figs. 11 and 12. This particular simulation laboratory consists of a spherical dome, 40 ft in diameter, attached to a two-story building, 40 ft square. In the center of the dome is a moving base pressure-tight gondola which duplicates the space vehicle crew's flight control station. It contains the instruments, controls and general cockpit arrangement of the space vehicle under investigation. The environment stresses of various space missions are reproduced therein: re-entry heat is produced by heating elements which can raise the inner wall temperature to 350°F; rocket noise up to 130 dB is produced by tapes and projected by a speaker system into the cabin; vibrations to 50 c/s and $\frac{1}{16}$ in. double amplitude are programmed as a function of flight condition by a signal generating device which drives the seats' vibrations bases; and the internal atmosphere in its principle elements of composition and mixture ratio, cabin pressure altitude, humidity and circulation rate is generated by systems of pumps, valves, supply bottles and regulators. The entire cabin environment is programmed as a function of the particular mission profile.

External vision is projected on the interior of the spherical dome by a planetarium type star field–Sun–Earth–Moon projection system, which is attached to the upper surface of the gondola. It is servo driven by signals from the computer. The Earth's and Moon's orbs may be varied in size from a pure horizon to spheres (including phasing) of the proper size when viewed from the vehicle's space position. Black light would be used to simulate the high contrast conditions of space.

The moving base has pivot points in pitch, roll and yaw at a point far enough behind the gondola to provide realistic short period translational and rotational motions. The entire gondola and gimbled structure is mounted on a yoke which pivots to provide the proper direction of the acceleration fields throughout the mission.

Atop the gondola is a pressure-tight hatch and tunnel which mates with a space cabin laboratory at the conclusion of the boost phase. This space cabin is a fixed base chamber, 8 ft wide and 25 ft long, which contains the same atmosphere and pressure as the gondola. When large vehicle long term space missions are being studied, the gondola and space cabin are used as a single vehicle, however, they may be used separately.

The entire simulator is driven by an analog digital computer complex capable of computing, in real time, the flight mechanics in six degrees of freedom, the navigation equations, rotating Earth, winds, programs for the environment generating systems, and such functions as required by the work stations in the space cabin. The equations of motion, aerodynamic and reaction control parameters, orbital mechanics, etc., are computed in a nearly total sense with only minor linearizations and compromises to simplify the apparatus. None of the parameters which might have a realizable effect on the crew would be omitted.

The master control station in the center of the laboratory consists of a large console which contains all the controls and monitoring devices (including closed circuit television) required to direct the simulated flights. Variplotters, pen and tape recorders record all the essential parameters for crew monitoring and crew proficiency evaluation on a continuous basis.

This entire facility is completely versatile in that it would be adaptable to any spacecraft configuration from a single-man orbital craft to

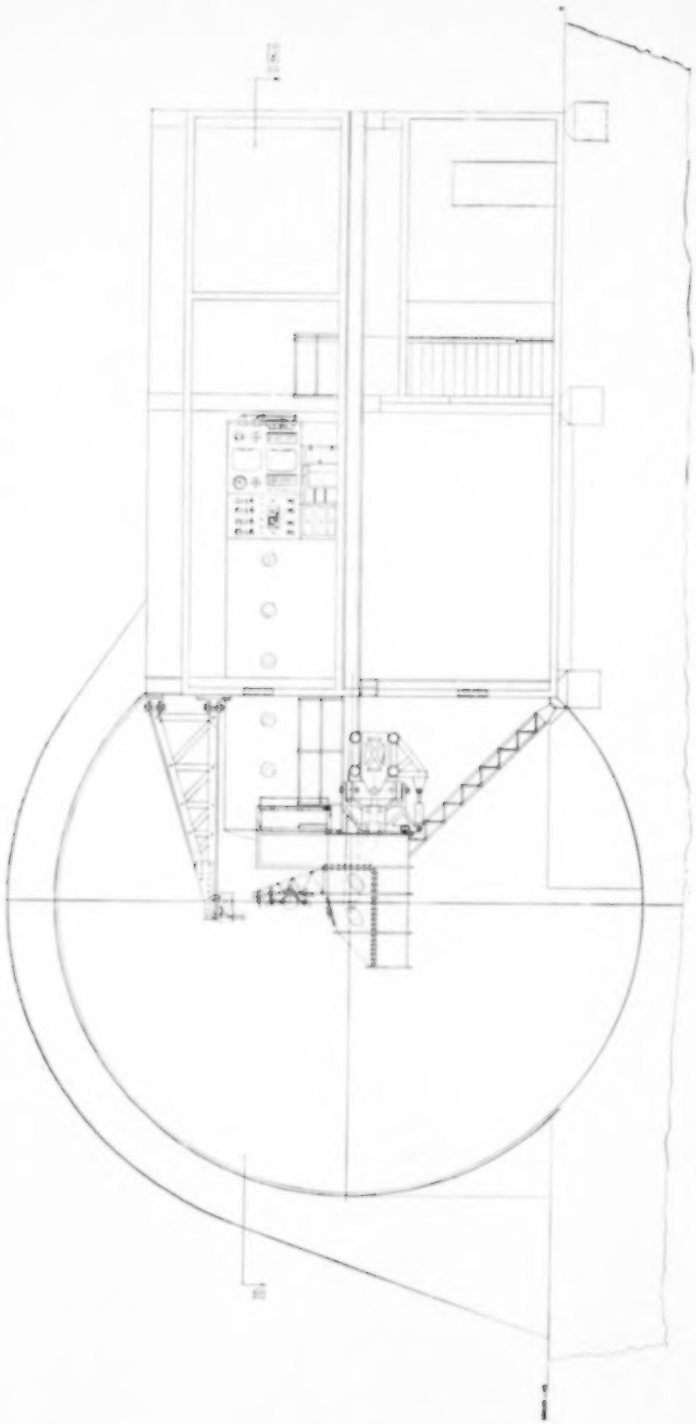


Fig. 11. Concept for space flight simulator laboratory (profile view).

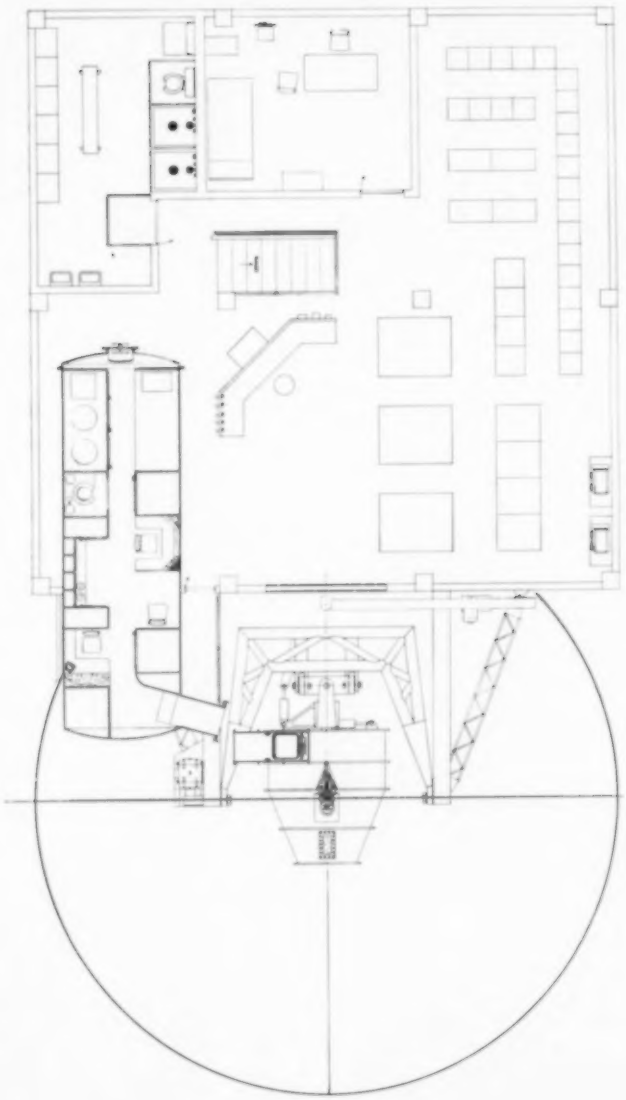


Fig. 12. Concept for space flight simulator laboratory (plan view, section through *BB*).

a three-man space voyaging craft, ballistic or lifting, and would be likewise adaptable to any mission within the capability of these craft. The computer is designed for universal application and the parameters and functions may be changed or reset to conform to the particular vehicle under study. The gondola and all apparatus associated with it are independent of the apparatus associated with the space cabin. Thus, the two units may be used together or separately.

No doubt there are other simulator concepts which could evaluate man's performance on space missions as well as the one described. However, to obtain valid data on man's performance on space missions, the simulator must have as many as practicable of the following basic parameters: combined stresses, vehicle dynamics, visual stimuli, moving base, and vehicle displays and controls. Although initial results indicate that man can assume a much more responsible role in space vehicles than previously thought possible, until such a total simulation program is implemented, man's true capability in an integrated man–machine space system can only be surmised.

SPACE FLIGHT ACCELERATION SIMULATOR

CHARLES L. BARKER, Jr.
Army Ballistic Missile Agency, Army Ordnance Missile Command

Abstract—This paper outlines the concepts of the Space Flight Acceleration Simulator, a ground-based training and research facility, which produces the acceleration-time history of rocket flight and the "g" fields of the Moon and Planets.

Its track captured capsule carries several trainees or hundreds of pounds of equipment through programmed "flights" with boost, zero-g and re-entry, or any combinations thereof.

Simulation of extraterrestrial gravitational and environmental conditions also allows full scale training and testing of personnel and equipment.

I. INTRODUCTION

With manned space flight just around the corner, consideration is being given to the final selection and training of personnel who will make these flights. One of the major problems arising in the training program is that of subjecting the astronaut to the realistic acceleration and weightless states which he will experience during his flight whether it be in a ballistic transportation missile or in an orbital space vehicle.

Many systems have been proposed and several schemes have been used to provide accelerations[1], short periods of zero-g[2] (weightlessness), or combinations of these. Each of these systems has its shortcomings. Studies have been made of the use of elevators in the Empire State Building and Carlsbad Caverns, and of the Gravitron, a U-shaped vertical track to establish zero-g conditions. Acceleration studies[3] have been conducted on centrifuges and tracks, and acceleration and sub-gravity has been accomplished with aircraft flights[4].

Aircraft flight is the only means now available (excluding for the moment the ballistic missiles) in which acceleration and sub-gravity conditions, which may reach 1 min in the near future, may be combined. The primary drawbacks in the aircraft flight programs, are the inability to obtain realistic acceleration and deceleration forces over the proper time periods and to provide a true weightless state during the full Keplerian flight. Additional problems connected with the use of aircraft are maintenance, weather, special aircraft and a relatively high cost per test.

Training programs will be conducted using ballistic missiles and in all probability with orbital vehicles; however, these programs are costly, limited in size, require recovery operations and actually, it would appear that the first flight might be equivalent to graduation.

Because of the limitations of the aircraft flight programs, there appears then to be a need for a ground-based facility capable of subjecting the trainee to the accelerations and weightless conditions as he will experience them in actual flight. To aid his training further, he should be provided with the capsule in which he will ride and be given the problems he must accomplish during his flight.

This paper describes the study of a facility which if built would accomplish these ends.

II. DISCUSSION

A. General

The Space Flight Acceleration Simulator is a ground-based, man-carrying training and research facility capable of accurately and repeatedly producing the acceleration and deceleration patterns of single and multi-stage rocket-boosted vehicles. Of major importance is the capability of producing true zero-g conditions up to 30 sec between the simulated injection and re-entry phases of the flight.

The unusual requirements of the facility require a unique design which in turn requires

that special consideration be given the location of the facility. Fig. 1 shows the facility as it might appear in the Grand Canyon.

Fig. 1.

B. The Facility

The Space Flight Acceleration Simulator consists of three principle elements. The first two of these, the track and the vehicle, will be discussed in detail later in this paper.

The third element would include the medical and physical research and test equipment and support laboratories shown at the base and the top of the track in Fig. 1.

The track consists of a horizontal circular track 1000 ft in diameter connected to a vertical track with a total height of about 1 mile.

The vehicle or capsule is secured to the rails and may travel several times around the circular track before and after being switched into the vertical or zero-g portion of the track. Propulsion is obtained by a low thrust turbojet with a swivelable nozzle for control purposes. Several capsule designs will provide vehicles for various training, materials checkout, and research missions.

Installation of lighting will provide round-the-clock operation.

C. Design Criteria

Many desirable and required features were listed during the conceptial design study of the acceleration simulator.

It was felt that one of the primary requirements would be that of a ground-based system. Since this suggested a track of some design, the

problems associated with track design could be kept to a minimum by having the track subsonic. Highly desirable also was the consideration that the sled or capsule be large enough to allow the trainees to be accompanied by a skilled observer who could exercise control throughout the test.

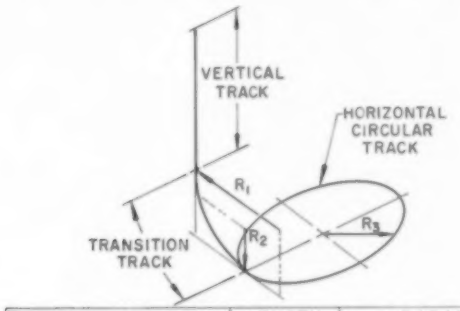
The capability to program any desired "flight" into the facility was also listed under the initial requirements as was the desire to be able to simulate the acceleration fields of extra-terrestrial bodies.

How well the design satisfies these requirements will be shown in the following paragraphs.

D. Track Configuration

Two different track configurations were initially considered in this study but the flexibility of the system shown in Fig. 2 eliminated further consideration of a fixed parabolically curved track.

As shown, this track has a horizontal circular section with a 500 ft radius. This is connected to the 3622 ft vertical track by a transition section which starts with a 500 ft radius and opens out to 1210 ft at the vertical track. A total track length of 8200 ft is required for this design.



	LENGTH	RADIUS
VERTICAL TRACK	3622	—
TRANSITION TRACK	1395	1210 — 500
CIRCULAR TRACK	3142	500
TOTAL HEIGHT	4519	

Fig. 2. Track layout.

The track, as seen in Fig. 3, consists of two heavy railroad rails set at 90° to each other. Since the "flight" will be conducted at subsonic speeds, special rails, continuous welding or accurate alignment may not be necessary.

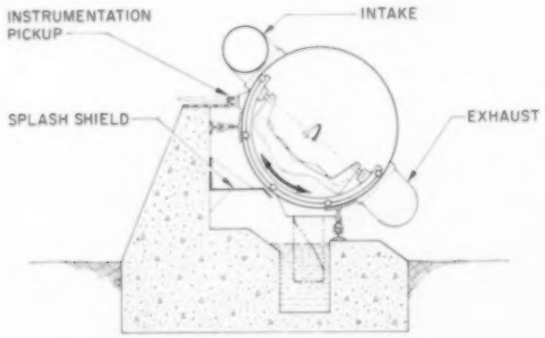


Fig. 3. Typical sled and track section.

The continuous water trough used to obtain deceleration forces and to provide emergency stop capabilities is shown in this figure. The water

splash shield and the instrumentation pickup are also shown.

The switch arrangement is shown in Fig. 4, indicating the method of removal of the wall and track sections. In this view, the lower rail of the transition track is dropped to the bottom of the water trough allowing continuous operation of the water brake if needed. Other sections are hinged or retracted as shown by the arrows to provide clearance for the capsule.

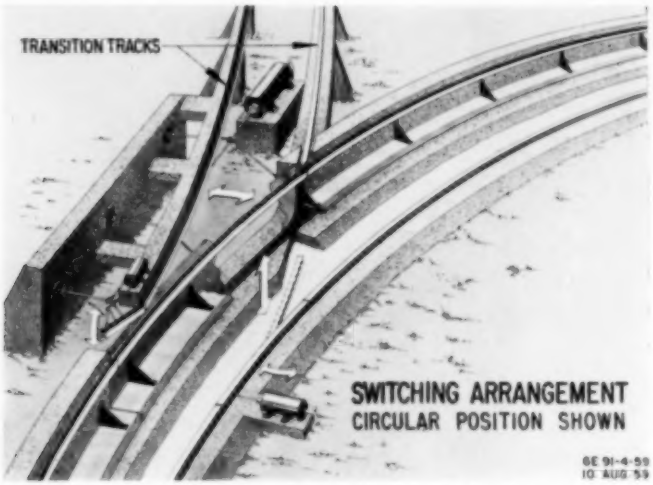


Fig. 4. Switching arrangement circular position shown.

Referring back to Fig. 3, one sees how the g forces are introduced to the trainee. A pendulum type or gimbled space bed insures that the resultant forces (Earth gravity, longitudinal acceleration or deceleration and centrifugal forces) are reduced to a single force.

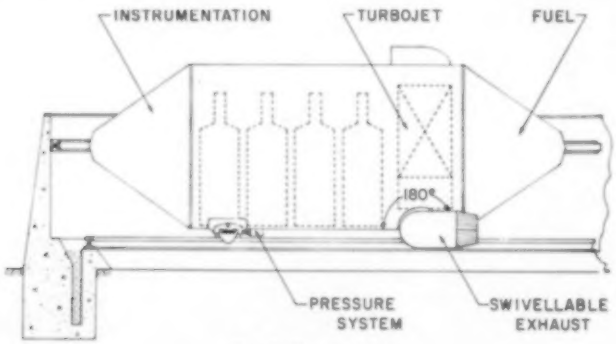


Fig. 5. Typical sled.

E. The Simulator Capsule

A typical manned training capsule is shown in Fig. 5. This 24 ft long capsule carries three trainees and an observer within its main compartment. The engine compartment houses a turbojet engine mounted radially to the vehicle axis while fuel and instrumentation are located in the streamlined end sections. Since the vehicle travels in both directions, the turbojet intake and exhaust ducts are swivelable.

An automatic safety system operable from within the sled or by external sources will stop the sled if necessary by extending the water brake or applying other devices in the transition curve and on the vertical track.

This two ton sled rides on shoes of special

design, required because of the varying curvature of the tracks. An interconnected pneumatic or hydraulic system on each shoe equalizes the loads on the several bearing pads. Since the water trough is continuous, water lubrication and cooling [5] of shoes may be accomplished by a small scoop.

Included in the cabin would be communications equipment (to the track control officer), a display showing "flight" instrumentation, a programmer to provide the desired "flight", emergency controls, etc. The space flight trainee or the research subject would perform certain operations during the flight under the viewing by TV or by the observer who may ride the sled.

Sound proofing of the cabin will be required and actual flight sounds and environmental conditions [6–9] may be introduced to provide realism.

Other designs of manned or unmanned sleds might be suggested here but it is not the intent of this report to describe in detail the equipment for the wide variety of missions which might use the facility.

An artist's conception of the capsule is shown in Fig. 6.

F. Propulsion

Because of the unique track design, the power requirements for typical missions on the space flight simulator are relatively small and simple. The requirements consist of a low power or sustainer type system used over the entire "flight" and a high power or boost phase required while on the transition track. Since the flight is programmed both ways on the track, the power supply must be capable of applying thrust in either direction as required.

Fig. 6.

Aerodynamic drag, frictional losses and longitudinal acceleration make up the three forces which establish the power requirements for the sustainer. These are calculated in the conventional manner and are combined to determine the total sustainer requirements.

It is found that the propulsion requirement for the sustainer system for the sled shown in Fig. 5 is in the order of 2000 h.p. or can be met with a thrust system developing slightly more than 3700 lb of thrust.

The boost or high thrust requirement may be accomplished by a jato with variable orientation as required, applied on entry into the transition

track. A thrust of slightly over 11,000 lb is required over this short period. Depending on the deceleration values desired in the flight, it may be necessary to apply an additional high thrust as the sled re-enters the transition track from the vertical.

As can be seen from a study of the 2000 h.p. required for the sustainer, the aerodynamic drag comprises approximately one half of this figure. Since drag is a function of the presented area, 4 m^2 in this case, any reduction of this area will result in a decrease in the power required to overcome the drag. Thus the sustainer for smaller sleds would be somewhat smaller.

The propulsion unit for the booster will be a solid or liquid rocket engine due to the high impulse short duration requirement. Many units are available and because of this, no further comment will be made other than the fact that the unit will weigh approximately 200 lb and will have thrust vector control obtained by either swivel nozzle, jet vanes, or tilting of the motor.

Several types of power plants can be used for the sustainer. Both turbojet and rocket engine appear to be very good and a turbo-prop might be used although reaction time and thrust directional control might present problems.

The use of a low thrust rocket engine appears to be very attractive until the propellant requirements for a "trip" of 5 min or longer are considered. Such a system would require approximately 19 lb of propellant/sec at maximum delivery rates. If a throttlable engine were used then the rocket engine might be first choice for its lightness and simplicity.

The low fuel consumption of the turbojet makes this engine appear very favorable also. Engine weight is much higher but the fuel consumption is about one tenth that of the rocket system.

For both of these systems, the thrust vector must be variable thru 180°. The small rocket engine might be completely rotated while a rotation of the exhaust cone of the turbojet would probably be the solution as shown in Fig. 3.

A more detailed study should be accomplished prior to the choice of any one system, to insure that all items such as initial cost, upkeep, operational costs, exhaust products, etc., are considered.

The turbojet was used for this study. Engine operation at fixed throttle with thrust directional control will provide the fast reaction time needed for close g tolerances.

Operation of the sustainer during the vertical or zero-g flight will require the thrust vector to be close to normal to the track. Only that vertical vector required to overcome drag and friction will be required at this time. A simple accelerometer will control the thrust direction, maintaining zero-g within the tolerance of the sensing device.

Propulsion units for the sled which may be considered are the J-33 (Allison) or J-34 (Westinghouse) turbojet engines, the HS-R-804 (Hispano-Suiza) engine or the Pratt and Whitney JT 12A-1 which has a dry weight of only 430 lb.

For programs with low performance, a wheeled sled with automotive propulsion and drive systems is entirely feasible.

III. SYSTEM KINEMATICS

The choice of the particular track configuration described above was based on many considerations. The requirement for sleds with cross sectional areas up to 4 m^2 was considered highly desirable as was the requirement for 6–8 g during boost and deceleration phases of flight. In view of the late data showing human endurance of 20 g, this acceleration value was also considered. To keep the power requirements low it was necessary to conduct the "flight" at as low a velocity as possible.

The circular track radius of 500 ft gives a 20 g capability with a velocity of about 400 m.p.h. as shown in Fig. 7. While smaller radii would induce an even greater g load at low velocities, any shorter "time-around-the-track" than that for the 500 ft radius (see Fig. 8), was considered to be too short to accomplish switching and still have a safety margin for stopping in the event of trouble. Since the power increase for larger radii track is a function of the velocity squared, the 500 ft radius appeared to be a good choice.

Final track design might result in other configurations for propulsion or design reasons.

Design of the vertical track was based on the desire to obtain at least 30 sec of zero-g. Fig. 8 also shows the height of the vertical track versus time at zero-g, thus 3622·5 ft of vertical track was required for 30 sec. Also shown on this figure is the velocity at which the sled must start in vertical "flight". For 30 sec, an entry velocity of 483 ft/sec is required.

Figure 9 shows the effect of the relatively low longitudinal accelerations when applied for extended periods of time. Since the circular track design takes advantage of centrifugal force as do the common centrifuges it becomes easy to produce high acceleration resultants with low power requirements. For example, the application of 0·08 *g* will result in a velocity of 300 ft/sec and a resultant acceleration of 6 *g* if applied for 120 sec. Fig. 10 describes the time and distance required to stop the sled at various *g* levels, given its velocity. This information completes that needed for switching and emergency track actions.

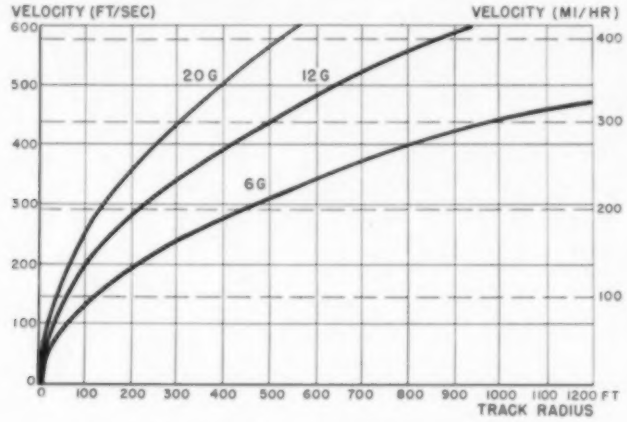


Fig. 7. Track curvature considerations.

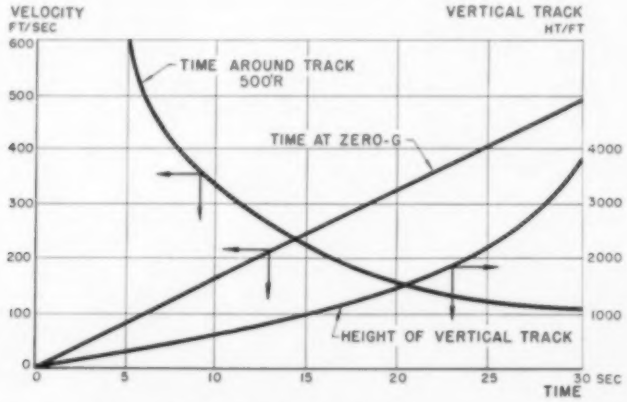


Fig. 8. Sled velocity parameters.

Since the use of a 500 ft radius for the transition section with the desired velocity of 483 ft/sec would result in a 14 *g* burnout acceleration, the transition curve opens from a 500 ft radius at the circular track to a 1210 ft radius at the start of the vertical track keeping the 6 *g* loading.

In performing a simulated flight, the desired acceleration-time history is prepared. A typical

"flight" of a two stage orbital vehicle with re-entry is shown in Fig. 11 and this is the acceleration-time pattern to be duplicated. As seen in Fig. 12, the boost and re-entry phases are identically matched, however, the zero-g or coast period has a limit of 30 sec (for this track). Note that velocity and distance traveled are also shown in Fig. 12.

Considering first, the first stage of the boost phase of the flight, application of $\frac{1}{10}$ g for 90 sec results in a velocity of 300 ft/sec at which time the water brake reduces the velocity to simulate first stage burnout. This maneuver, without braking, is repeated for the second stage after which the sled enters the transition track.

To take advantage of the maximum weightless

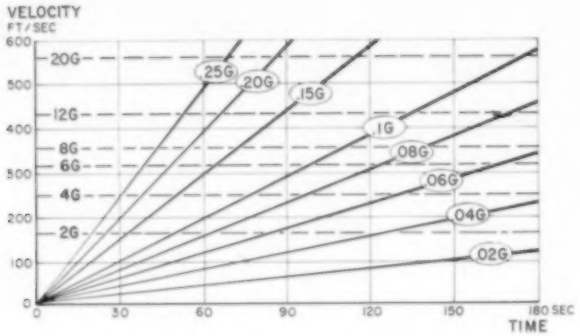


Fig. 9. Longitudinal acceleration study.

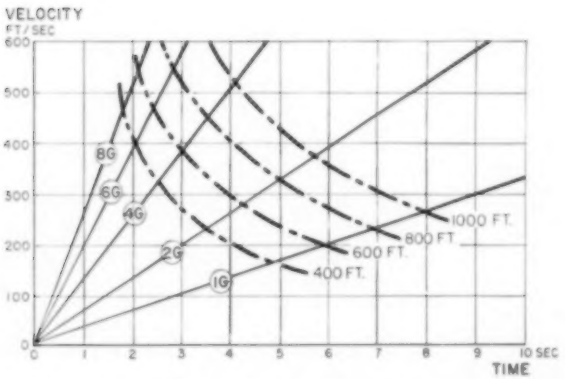


Fig. 10. Sled deceleration considerations.

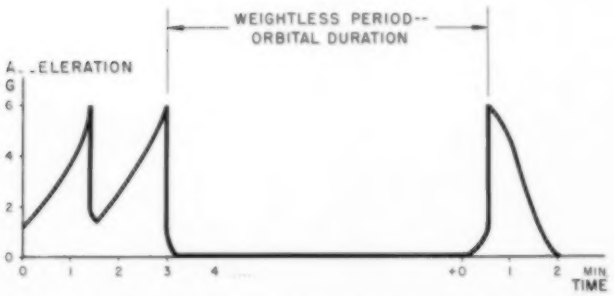


Fig. 11. Typical acceleration pattern two-stage booster.

time available, the velocity is increased during the travel on the transition track as indicated. This velocity increase over this short distance requires a jato assist. Since the entry into the vertical flight is tangential to the transition track, the acceleration due to centrifugal force drops to zero. Only a small fraction of the propulsion available is required during the upward and downward portions of the flight to balance

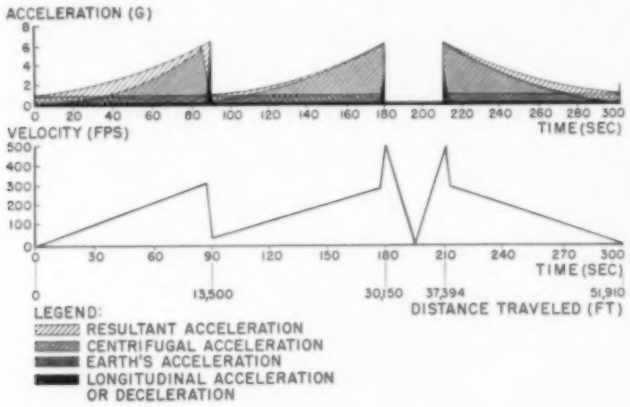


Fig. 12. Typical two-stage boosted capsule.

the aerodynamic drag and friction forces tending to hold back the vehicle. Since the control of these external forces may be held to the accuracy of the present missile control systems, there appears to be the possibility of obtaining zero-*g* or any desired level of sub-gravity with a tolerance of $\pm \frac{1}{100}$ of a *g*.

Upon re-entry into the transition track, the vehicle is again subjected to the desired *g* forces. The simulation of opening of parachutes and ground impact are accomplished by high-impulse short duration rockets. In this simulated flight, the sled has traveled a total distance of over 10 miles in 5 min.

An increase in time on the vertical track is possible with values of sub-gravity other than zero-*g*. On the 30 sec zero-*g* track, the Martian gravity of 0·39 Earth *g* may be held for 39 sec while 0·7 Earth *g* can be obtained for 55 sec.

Other programming is possible such as a boost phase followed by a zero-*g* flight to the top of the track where the sled is stopped. It also might be desirable to start with zero-*g* and perform a re-entry maneuver. These partial flights would simulate the conditions which would be experienced by personnel going to orbital stations for extended time periods or returning from there to these stations.

With the application of proper maneuvers, it is possible to enter the zero-*g* flight with accelerations in excess of 6 *g* by accomplishing a braking maneuver just prior to vertical flight. Automatic orientation of the passenger allows him to feel the resultant forces only, so he does not recognize the braking maneuver as such. This braking maneuver is required to bring the velocity down to 483 ft/sec since excess velocity would cause an over-run at the top of the track. A study of this high pre-weightless *g* followed by weightlessness may allow the researcher to raise the presently planned burnout acceleration limits for manned flight to more efficient levels. Using the circular track only, accelerations of over 70 *g* can be obtained with subsonic velocities.

The above cited examples illustrate the flexibility of the simulator to provide time-acceleration curves for any desired "flight" conditions and to simulate lunar and planetary gravity fields for test of equipment and personnel under these conditions.

From this *one* facility it is possible to study the biological effects of zero-*g* and sub-gravity, sustained and impact accelerations, angular,

oscillatory and jostling accelerations and of course the accelerations of simulated flights as these effects influence plant and animal life.

IV. SAFETY

One of the most obvious advantages of the space flight simulator is the ability to conduct entire flight programs with extreme safety in the ground environment. The fact that gradual controlled build up to final accelerations may be conducted over a series of flights with a skilled observer aboard adds confidence and may materially increase the number of acceptable space flight personnel.

V. CONCLUSIONS

The following conclusions may be drawn from the foregoing work:

(1) No system now exists which will provide true zero-g conditions for periods of 15 sec following and followed by accelerations typical of rocket powered flight except a rocket flight itself;

(2) it is possible to provide such a man-carrying system with a ground-based, low velocity track of unique design;

(3) as the system is track captured, it provides a safe training and research tool for future space flight programs;

(4) flexibility in programming the "flight" allows the matching of any desired rocket trajectory as well as a gradual build-up to the full load conditions for trainee programs and, finally,

(5) it appears that the system may be built for less than the cost of a single ballistic missile training device.

REFERENCES

1. J. P. Stapp, *Astronautics*, Vol. 4, No. 2, (1959).
2. H. von Beckh, *Astronautics*, Vol. 4, No. 2, (1959).
3. *Possible Methods of Producing the Gravity-free State for Medical Research*, Proj. Rep. USAF School of Aviation Medicine (Jan. 1951).
4. S. J. Gerathewohl, O. L. Ritter and H. D. Stallings, Jr., *Producing The Weightless State in Jet Aircraft*, 8th International Congress, Barcelona (1957).
5. M. R. Walfson, *Shoe Design and Materials Studies for Hi-Speed Sliding Friction Applications at Snort*, NOTS Rep. No. 1845, Navord Report No. 5626.
6. C. C. Clark, J. D. Hardy, *Astronautics*, Vol. 4, No. 2, (1959).
7. E. P. Hiatt, *Astronautics*, Vol. 4, No. 2, (1959).
8. K. E. Schaefer, *Astronautics*, Vol. 4, No. 2, (1959).
9. E. B. Konecci, *Astronautics*, Vol. 4, No. 2, (1959).

CLOSED ECOLOGY PROBLEMS IN A SPACE CABIN SIMULATOR

JEAN R. NELSON

Minneapolis-Honeywell Regulator Company, Aeronautical Division

Abstract—A two-man space cabin simulator for research and training is being designed and built by Minneapolis-Honeywell for the Air Force School of Aviation Medicine. All provisions for physiological functions are to be completely contained within the sealed cabin for 30 days. Two major problems have been (1) reprocessing of urine and waste water for re-use and (2) complete control of atmospheric content. A small reprocessing plant has been fabricated with provides a usable water yield ranging upward from a minimum of 60 per cent depending on the amount of urine dilution. Cabin atmosphere is automatically controlled to the desired temperature, humidity, and partial pressures of oxygen, carbon dioxide and filler gas, utilizing a complex array of sensors and control techniques.

Concern for man's well-being in the man–machine system has become a practical necessity in recent years. One of the well recognized problems to solve in assuring man's well-being on the stairway to space is his adaptability to capsulated ecology. The Space Medicine Division of the School of Aviation Medicine recognized early the need of research on closed ecology of a space cabin and plans for a simulator were initiated by Dr. Hubertus Strughold and Dr. Fritz Haber in 1952. Under the direction of George R. Steinkamp, Lt.-Col. USAF(MC), Chief of Astroecology Department, a series of five 7-day flights have been accomplished in a one-man 96 ft^3 space cabin simulator.[1] These flights, starting in February 1958, were preceded by several years development of equipment and techniques.

At Minneapolis-Honeywell, we were invited last fall to share in the expansion of this research program by designing and constructing a two-man, 380 ft^3 space cabin simulator for the School of Aviation Medicine. The space cabin simulator is primarily a research tool, although its use may be eventually expanded to aid in the selection of astronauts and ultimately to serve as a trainer.

Since the $5 \times 8 \times 12$ ft elliptical cylinder forming the shell of the cabin is completely sealed for 30-day flights, provisions for all physiological requirements such as breathing, eating, drinking and personal hygiene must be completely contained within the cabin. Herein lay the challenge to our designers in developing this research vehicle.

Many minor problems kept life interesting for the designers, but the two areas of atmosphere control and waste conversion, discussed in this paper, proved most worthy in challenging their ingenuity. Neither of these problems was solved with a startling breakthrough in state of the art, but rather, a practical combination and expansion of known processes was accomplished.

Cabin atmosphere is automatically controlled to the desired values of temperature, relative humidity, and the partial pressures of oxygen, carbon dioxide and the filler gas. Carbon monoxide is also sensed and controlled through use of a catalyser.

Interactions between pressures, temperature and relative humidity control systems occur so an analog simulation study was conducted.[2] There were no detrimental interactions that affected stability of any control system. The working design requirements established by the School of Aviation Medicine and Minneapolis-Honeywell for atmosphere control are listed in Table 1.

Table I. Working Design Requirements for Atmosphere Control

	Range	Threshold
Oxygen (partial pressure)	100–400 mm Hg	± 4·5 mm Hg
Carbon dioxide p. p.	0–80 mm Hg	± 1·3 mm Hg
Carbon monoxide p. p.	0–0·07 mm Hg	0·07 max.
Total pressure	240–760 mm Hg	± 4 mm Hg
Altitude equivalent	28,500 ft–SL	± 500 ft
Pressure leakage	0–0·052 mm Hg/hr	0·052 max.
Mean temperature	60°–80° F	± 1°
Relative humidity	40–60%	± 3 at 40%

Location of the physical plant for handling the cabin air is shown in photograph, Fig. 1. Cabin air is collected through floor register openings in ducts running the cabin length on both sides and collected in a plenum at the hatch end of the simulator. The electrostatic air cleaner is located between this plenum header and the circulation blower, which moves air toward the console through the underfloor ducting. The carbon monoxide catalyser is mounted so that blower discharge pressure forces air through the catalyser to the return ducts. The blower discharge duct is further split so that up to one-third the total air flow can be diverted through chemical beds for adsorption of carbon dioxide. Distribution baffles downstream of the carbon dioxide adsorption beds direct all air uniformly through activated carbon beds for odor adsorption. The adsorption bed outlet plenum chamber serves a vertical air duct behind the instrumentation or work console. Air flowing upward is relieved of excess moisture in the dehumidifying coils, reheated with resistance heaters to the set cabin temperature, and teed into

Fig. 1. Air distribution system.

perforated distribution ducts along each side of the cabin about shoulder level. Low duct face velocities of about 25 ft/min prevent objectionable drafts.

Absolute pressure, temperature, relative humidity and industrial type gas sensor electrical outputs are used for interior and exterior console instrument display, strip chart recorder and operation of the control systems. Set point control adjustments and instruments calibration equipment are provided on both the interior and exterior consoles. The exterior or research monitor console retains override authority at all times whether it serves to establish research experimental conditions or to correct pilot error.

OXYGEN CONTROL SYSTEM

The liquid oxygen gaging system is shown in Fig. 2. Oxygen requirements for two men for 30 days are more than supplied by the two

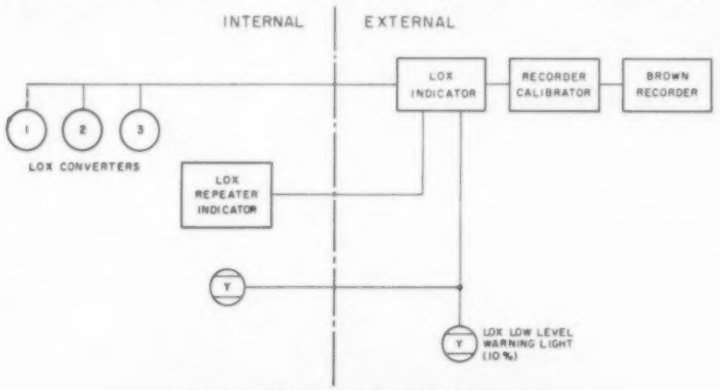


Fig. 2. LOX gaging system block diagram.

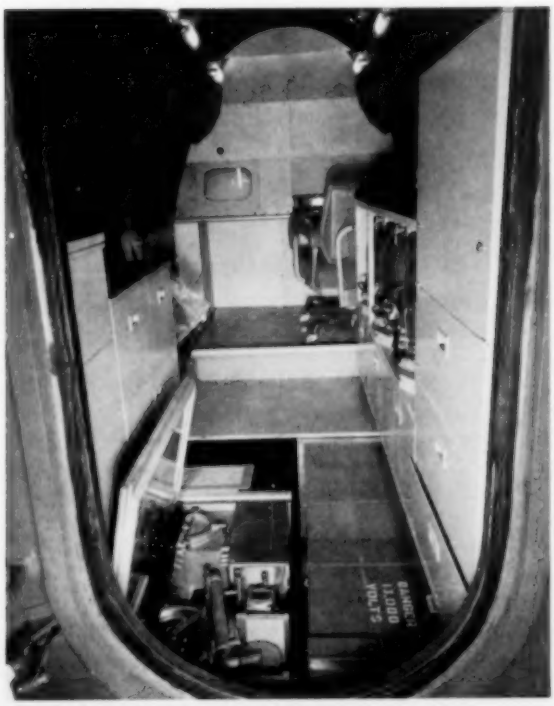


Fig. 3. Interior view of space cabin simulator.

25 l. LOX converters. The minimum boil-off rate of the LOX converters is presently greater than two-man usage of oxygen. Addition of a third LOX converter to supply oxygen for the organic waste incinerator, later described, provided flexibility of oxygen system operation without excessive overboard loss of oxygen.

The three converters are mounted in the right cabinet shown in Fig. 3. The converter pressure varies from 50 to 345 psi and cabin injection pressure is regulated to 50 psi. By scheduling waste disposal operation during higher converter pressure periods, little or no oxygen will be vented overboard.

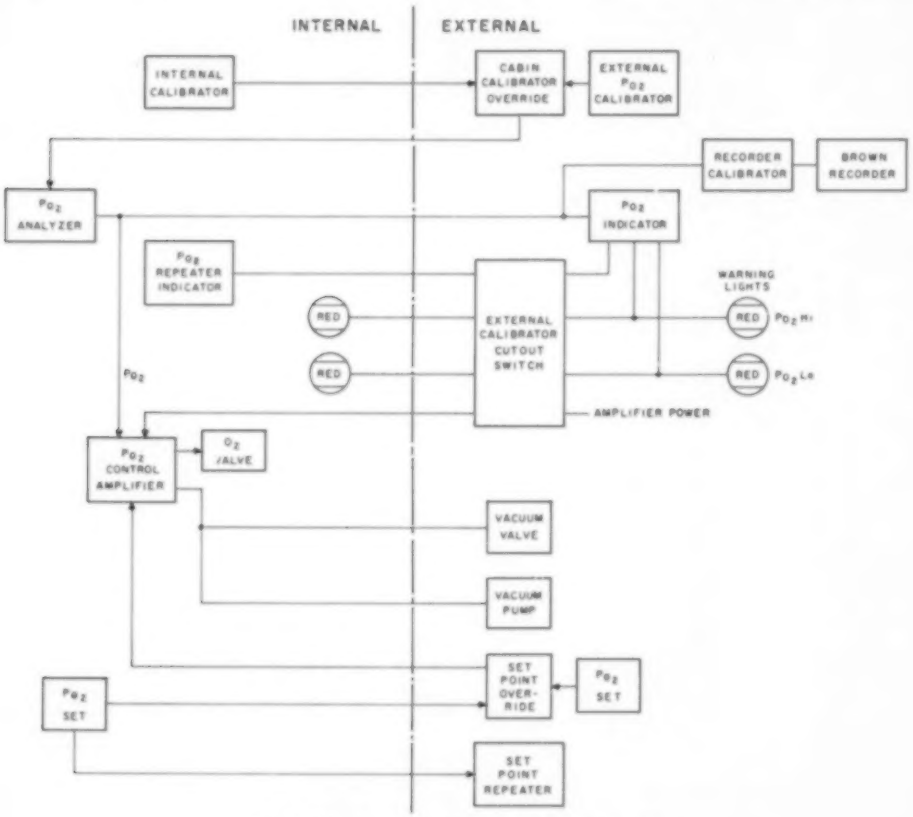


Fig. 4. Oxygen control system block diagram.

The oxygen control system is shown in Fig. 4. The d.c. output of the oxygen analyser is chopped, amplified and used to operate phase-sensitive relays. One relay operates a solenoid valve, which injects oxygen into the cabin if the error signal is less than set value, and the other relay operates the total pressure trim pump, which evacuates air from the cabin if the error signal is above the set value. The analog simulation showed no interaction problems between the oxygen and total pressure, temperature and relative humidity control systems.[2]

Calibration of the oxygen analyser is accomplished from inside or outside the cabin by flowing 150–200 cm^3/min of O_2, CO_2 and CO for the up-scale calibration and nitrogen for the zero scale reading. Research monitors outside the cabin can calibrate the analyser or override the interior set point control without the knowledge of the pilots.

CARBON-DIOXIDE CONTROL SYSTEM

The carbon-dioxide level will be maintained at the desired set level by damper control of the

amount of air flow through chemical carbon dioxide absorbing beds (see Fig. 5). The d.c. output of the carbon dioxide analyser is chopped, amplified and used to drive a phase-sensitive motor to actuate the damper bypass valve. The adsorption bed was divided into nineteen separate containers, each weighing about 7·75 lb empty. Sectioning the bed provided small air plenums through the bed which prevent channeling and consequent efficiency loss and reduced chemical handling problems.

The flow of calibration gas to the carbon dioxide analyser is in series with the oxygen analyser. This system, as in the oxygen system, provides override authority of calibration and set point values by the outside research monitors.

An auxiliary carbon dioxide system is provided as an emergency standby if emergency repairs are required to the main blower system or if additional absorbing capacity is required during experimental high-level carbon dioxide test conditions. This system has about the same rate of adsorption as the main system but much less capacity.

CARBON-MONOXIDE SYSTEM

The carbon-monoxide content of the air is reduced by catalytically converting carbon monoxide to carbon dioxide at 350°F. The thermostat and heater controlling the catalyser temperature are energized through a manually operated switch when the carbon monoxide analyser exceeds the maximum set point and lights a warning. Calibration of the carbon monoxide analyser requires 100 times the gas flow of the oxygen or carbon dioxide analysers, or about 15 l./min, so a separate flowmeter is provided to accomplish the calibration.

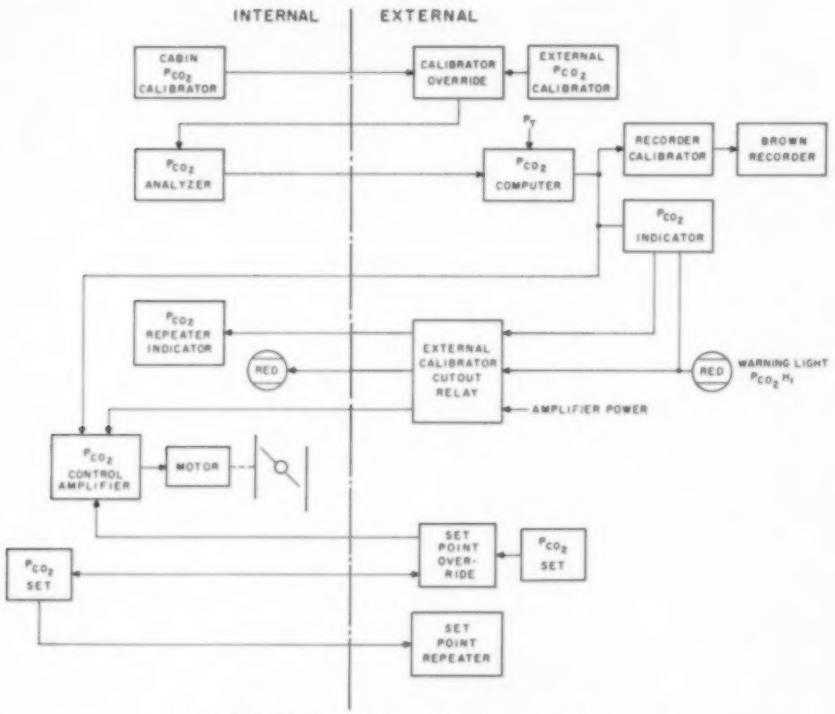


Fig. 5. Carbon dioxide control system block diagram.

TOTAL PRESSURE CONTROL SYSTEM

The total pressure control system is shown in Fig. 6. A water-sealed vacuum pump, capable of evacuating the cabin to the equivalent altitude of 28,000 ft in 3½ min is used for large decreases

in pressure. A smaller total pressure trim vacuum pump is used during most of the control regime. The total pressure transducer signal is amplified and operates phase-sensitive relays to energize a solenoid valve for injection of nitrogen to raise cabin pressure and the trim vacuum pump to lower cabin pressure.

ELECTROSTATIC AIR CLEANER

An electrostatic air cleaner was tailored for this application to trap dust and smoke particles (see Fig. 7). We felt that the electrostatic filter accomplished three important tasks in the air conditioning system: (1) The filter removes particulate matter such as lint, dust, dried mucus and skin particles, bacteria, and a good many viruses which could constitute a health hazard.

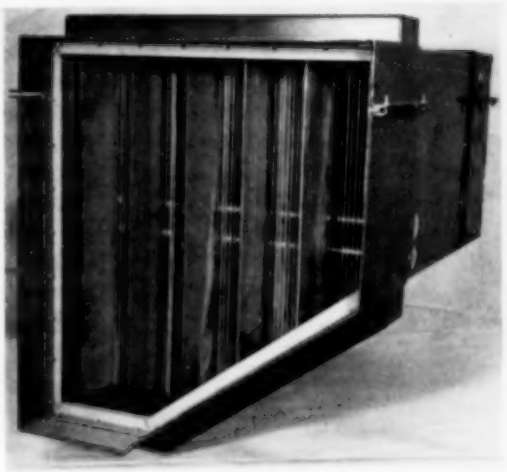


Fig. 7. Electrostatic air cleaner.

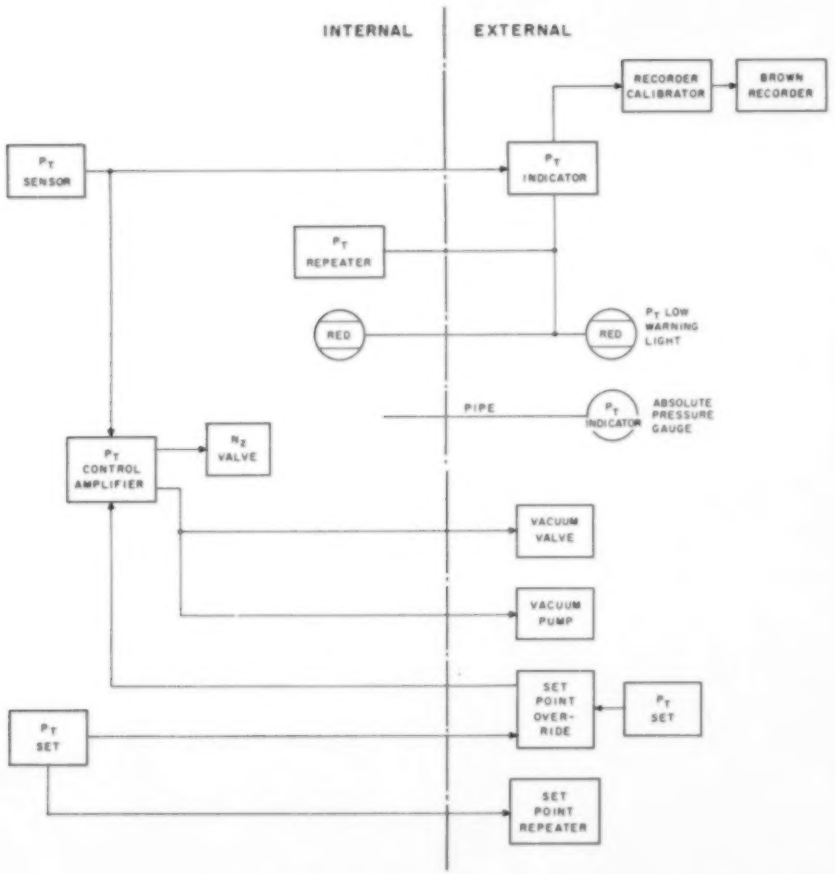


Fig. 6. Total pressure control system block diagram.

(2) The filter removes tobacco smoke, which roughly encompasses a particulate size range from 0·002 to 1·0 μ (see Fig. 8). We feel this feature of permitting tobacco smoking during a 30-day test will prevent psychophysiological bias of test results on tobacco-deprived pilots. This should expand the field of potential pilots. (3) Removal of particulate matter from the air stream will prevent clogging of the activated carbon filters. Air line experimentation with activated carbon beds for odor removal has shown rapid clogging of the carbon beds due to the inefficiency of the mechanical filters in removing particles from the air stream.

TEMPERATURE AND HUMIDITY CONTROL SYSTEMS

The close interrelation of the temperature and humidity control systems leads to discussion of their configuration together. Psychrometric charts for the normal and maximum equivalent altitude operating points (18,000 and 28,500 ft respectively) were computed for the gas constituencies to be used.

Early in the analysis of the humidity system, it was determined that the dominant process is the increase in relative humidity due to the latent heat load from the occupants. To increase the relative humidity from 40 to 60 per cent about 760 grains of moisture are required. Two pilots will add moisture at a rate of about 2000 grains/hr which would require about 23 min to span the total desired range of adjustment. Elimination of the need for adding water as steam or spray simplified the control system. There are, therefore, only two independent variables for controlling the cabin temperature and humidity: heater input and cooling coil brine temperature. It is noted at this point that heat transfer is through a chilled brine (15° to 40°F) system to prevent possible toxic refrigerant gas leakage into the cabin.

In sizing the circulation blower, cooling coil and heater commensurate with desirable draft levels and duct size, it was apparent early in the analysis that we did not have enough room to completely contain the temperature and humidity systems in the cabin. Since another requirement of the program was to provide a sound insulated room around the cabin to prevent extraneous diverting sounds, a dual function was conceived, whereby the walls of the cabin carried the steady-state heat load (see Fig. 9). Sizing the cabin equipment then required only capacity for minor transients.

The heat load of the cabin is distributed so that the console instrument heat load is handled by a separate thermostatically controlled cooling coil, the waste processing equipment has its separate heating and cooling control system in an insulated box, the steady-state coil and heater in the insulated room surrounding the cabin, and the cabin transients due to change in set point, cooking or other load changes are carried by the dehumidifying and cooling coil and heaters in the main circulation vertical duct.

The temperature control system is dominant over the humidity control. Normal operation will find the humidity control operating the cooling coil and the temperature control operating the heaters. If the cabin temperature increases, the temperature control reduces the heater input to the cabin. This continues until the heat is completely shut down. If the temperature continues to increase and the error reaches, say, 2°F, the no-heat limit switch will transfer authority over the cooling coil from the humidity control system to the temperature control system. During this operating regime, the humidity is not controlled. When the temperature error is removed, the first temperature decrease requiring heat addition switches the cooling coil operation back to the humidity control system.

There are three other override functions in the temperature and humidity control systems. One relay reduces the heater setting if incoming air exceeds a temperature of 85°F, and a second relay increases the heater setting if incoming air drops below about 55°F. The third override is thermostatic control of the brine temperature in the dehumidifying coil to 31°F for prevention of coil icing.

The set point for temperature control of the insulated room, which surrounds the cabin, is

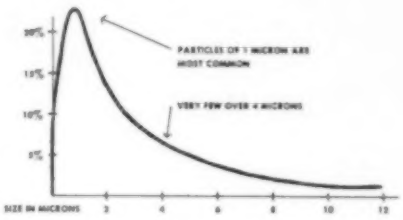
maintained approximately 5° below that of the cabin to carry the steady-state heat load of the cabin through its shell. The Honeywell home conditioning Modulflo principle is applied in this situation except in reverse. This is one of the few instances where we can control the outside weather, a weatherman's dream, from desired inside conditions.

This is how it works. Assume that a temperature transient from a set point change or operating condition change occurs that exceeds an error of, say, 1°F. The error is amplified to operate a proportional motor, so that the insulated room temperature bridge is thus proportionally changed with the amount of temperature error in the cabin, and the necessary action is accomplished by the room cooling coil or heater to help remove the cabin temperature error. The ratio of this system is about 5 : 1, that is, a 4°F transient error will result in a corrective change of about 20°F in the insulated room.

Turning now to the second area of interest in this program, let us look at a hypothetical situation: you are cooped up in your house for 30 days. No outside contact can be made, your children cannot carry out the trash accumulation, the garbage disposal is indisposed, your water is literally cut off, and the septic system is stopped up completely. Ugh! Don't think about it too long! These were the sanitation problems that were solved in the design and construction of the space cabin simulator.

The sanitation system for the space cabin simulator, shown in Fig. 3 (lower left) is compacted into approximately 18 × 30 in. of floor space and 18 in. below floor level. All organic waste such as paper, whiskers, hair, fingernail clippings, fecal matter and food scraps are incinerated in the waste disposal unit. All waste water such as condensate from the air conditioning coil, lavatory wash water, urine and the waste disposal scrubber tank output, which contains water of combustion and other dissolved chemicals, is processed to supply pure water for washing and drinking. This water would pass any health purity test other than psychological.

Let us look at these systems in more detail.



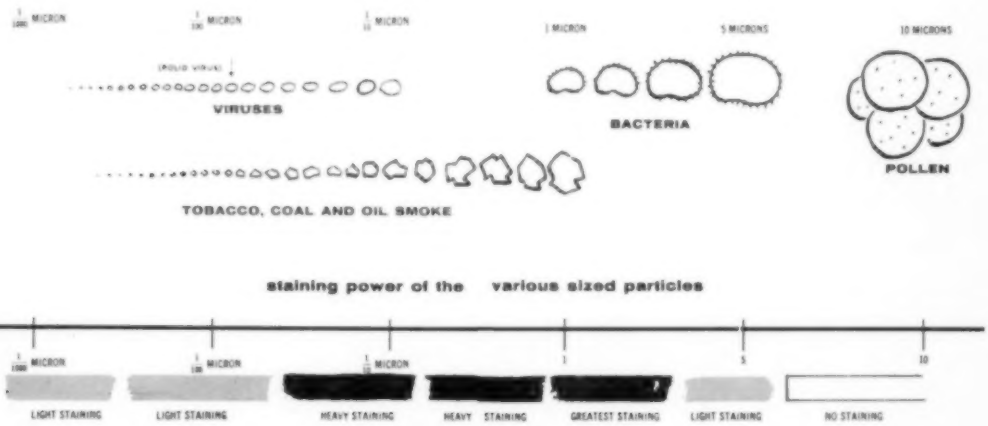


Fig. 8. Typical air pollutants.

Organic Waste Disposal System

The original concept for disposing of organic waste was to complete combustion within a

calorimetric type bomb or pressure vessel containing an excess of oxygen in the form of a chlorate candle or perchlorate of some form. This approach was abandoned quite early due to several problems such as the nuisance of mandatory dehydration before combustion to prevent excessive pressures, metal vessel pitting and handling problems with oxidizing chemicals. through the system to provide an excess of oxygen for complete combustion. The entire cycle is automatically time-sequenced.

Water Processing System

The complex chemical composition of urine led to its use as the sole criteria of success in achieving usable water from waste fluid resulting

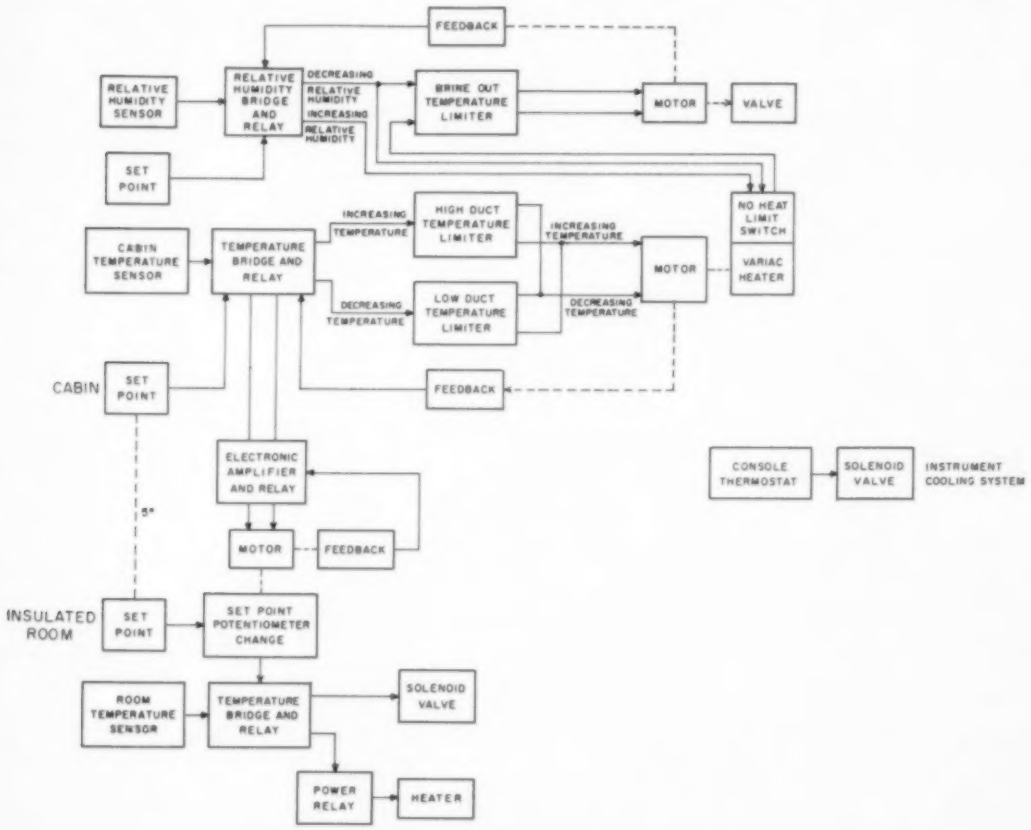


Fig. 9. Temperature and humidity control system block diagram.

The actual solution to this problem evolved naturally in the development program. Dehydration was already found necessary, and what better way to handle gasses from this process than with a high-temperature afterburner. A natural treatment to cool afterburner gasses and remove water is with a brine-cooled water scrubber. After the dehydration period, the burner chamber is raised to combustion temperature and a flow of cabin air pumped from cabin operation. The initial goal was to obtain water suitable for purposes other than drinking, for example, washing. This goal was exceeded by obtaining potable water. Tests have been successfully completed with all fluid waste products likely to be generated during cabin operation.

Referring to Fig. 10, the lavatory waste water and urine are collected in a cooled pretreatment storage tank. After treatment with sulphuric

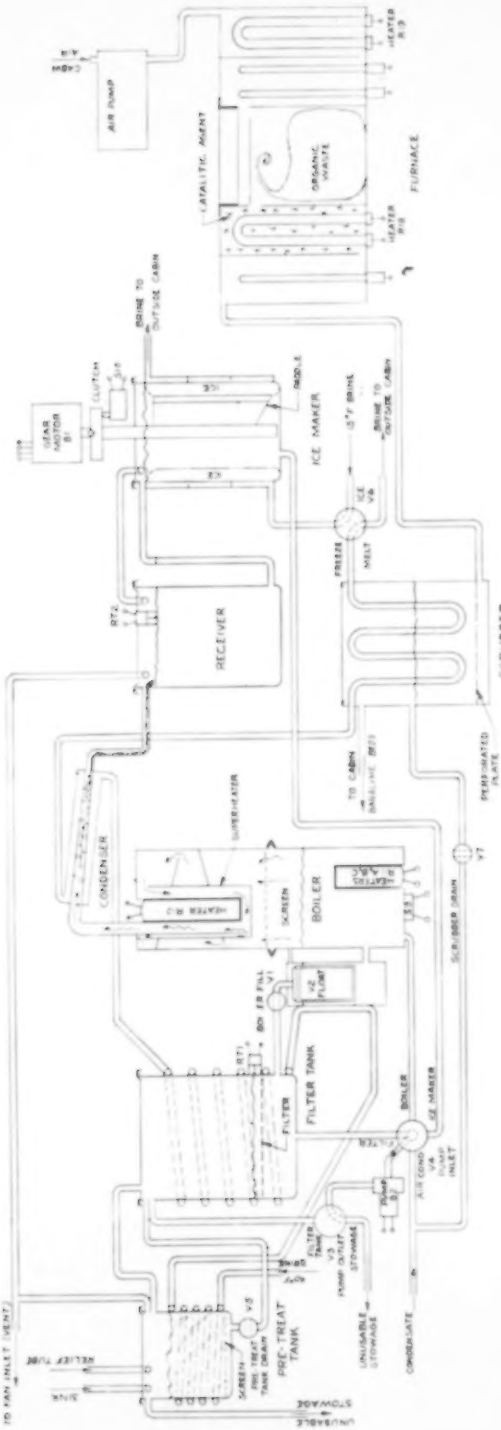


Fig. 10. Sanitation system flow diagram.

acid and anti-foam agent, it is gravity fed into the filter tank by elevating the pretreatment tank. The filter tank incorporates a mechanical filter and is the reservoir that maintains the desired fluid level in the boiler through action of a float valve. A thermistor level indicator shuts off the boiler heating elements if the liquid level drops too low. Vapor leaving the boiler is superheated for flash sterilization before it enters the condenser. The condensate receiver tank also incorporates a thermistor level override to de-energize the boiler heaters. Manual elevation of the prefreezer storage tank charges the ice-maker. The double-jacketed brine-chilled ice maker contains a paddle, which is slowly rotated by a small motor during the freezing process. When the hollow ice cylinder builds up to the desired thickness, the increased torque on the paddle operates a clutch device and actuates a microswitch. The microswitch turns off the motor and lights a console indication of ice completion. The ice is flushed and placed in the top of the storage tank. As the ice melts, it drips through a filter of activated carbon.

Periodic draining of the boiler is required to prevent concentration of impurities from "bumping" over into the condensate. A built-in safety factor is visual inspection of the ice. If the ice is hard and clear, it will contain something in the order of 30–50 parts per million of impurities when flushed and melted. If the impurities are in excess of the acceptable amounts, the ice will be porous like a popsicle and it acts as a sponge to soak up water with high concentration of impurities. This "popsicle" ice alone is satisfactory if centrifuged.

The work accomplished on this space cabin simulator sanitation facility has certainly satisfied the physiological and psychological requirements for this research vehicle, but certainly other requirements of a system for an operating space vehicle such as high to zero gravity operation and lightweight construction must soon be achieved. The subject of a future paper might well be a zero-gravity sanitation plant.

CONCLUSION

The space cabin simulator described in this paper is but one of several steps in the challenging process of putting man in space. In addition to valuable research into the psychological–physiological problems of space flight, the simulator program has developed many sensing and control techniques and equipment.

Continued development will extend and modify these techniques for use in future space vehicles.

REFERENCES

1. G. R. Steinkamp, W. R. Hawkins, G. T. Hauty, R. R. Burwell and J. E. Ward, *Human Experimentation in the Space Cabin Simulator*,
2. R. D. Schmidt, MH Aero Report 2249-TR1.

PACKAGE EXPERIMENTS FOR BIOASTRONAUTICS RESEARCH

LAUREL VAN DER WAL
Space Technology Laboratories, Inc.

Abstract—An argument is presented for increasing the pace and scope of flight-test experiments which will contribute to the earliest achievement of successful manned space flight. The use of "package experiments"—minimum-interference experiments which do not compromise the primary test objectives of the carrier vehicle—is proposed as a means of attaining this goal, thus extending an accepted missile development test philosophy into the field of bioastronautics research.

As examples of the type of research that can be accomplished, several specific flight-test experiments which can and should be performed in the immediate future are described. These experiments include component equipment testing and measurement of environmental parameters as well as the use of biological subjects to obtain physiological data. Weight, volume and power requirements are given for these experiments, and suitable vehicle trajectory parameters are described.

I. INTRODUCTION

In these days of primary concern with rapid development of effective weapon systems, it is easy to overlook the potentialities for experimental research in rocket-powered vehicles in areas other than improvement of vehicle performance. In fact, there is a long and proud history of important contributions to scientific knowledge and engineering skill through use of what may be called "package experiments", i.e. experiments which introduce minimum interference with the primary objectives of the rocket flight, and do not in any way compromise those objectives. Outstanding examples of such experiments are the discovery of the Van Allen radiation belts on the flights of Explorer I and Explorer III, and the successful maintenance of life in an artificial atmosphere for several days in space in Sputnik II.

The history of flight-testing in bioastronautics research includes three general categories of test philosophy:

(1) *Prime Objective*: In this case, the flight is designed primarily to obtain bioastronautics information. The Aerobee flights (1949–52) had as prime objectives the testing of an ejection escape mechanism and the study of the biological effects of the zero gravity state. The current Discoverer flight series was initially proposed to test vehicle performance, with the bioastronautics research program as a secondary objective. However, the recognized need for bioastronautics data has altered the purpose of this program to the point that the study of the physiological responses of four mice flown in Discoverer III were given equal importance with vehicle performance objectives; the mouse program is now a countdown hold item. The Mercury program, including all its development test flights, is another example; the prime objective will be to examine the response of man and his equipment to vehicle and space environment.

(2) *Secondary Objective*: These experiments are essentially "payload fillers," flown on vehicles whose prime objective is achievement of some vehicle-performance goal. In such cases, the payload capability is always less than adequate to accommodate all the proposed experiments; selection must be based on relative importance of the anticipated results as well as compatibility of the experiment with the vehicle mission. All of the experiments aboard the Pioneer flights were of this

class. It is probable that a recently announced Russian animal flight (and by inference all of the others) was in this category, since it was announced that two dogs and a rabbit were carried on a test flight of a new single-stage rocket "much more powerful than any yet tested by the U.S."

(3) *Hitchhike Experiments*: These experiments constitute late additions to the vehicle system, replacing ballast, for example, or mounted in one of the jettisoned stages rather than in the payload compartment. They require little or no modification of the telemetry or other vehicle subsystems to obtain the desired data. A recent example of this class of experiment was Project MIA, in which mice were carried in Thor–Able nose cones (Fig. 1). Heartbeat signals were transmitted over an unused segment of a commutated reference channel in the vehicle telemetry system.[1]

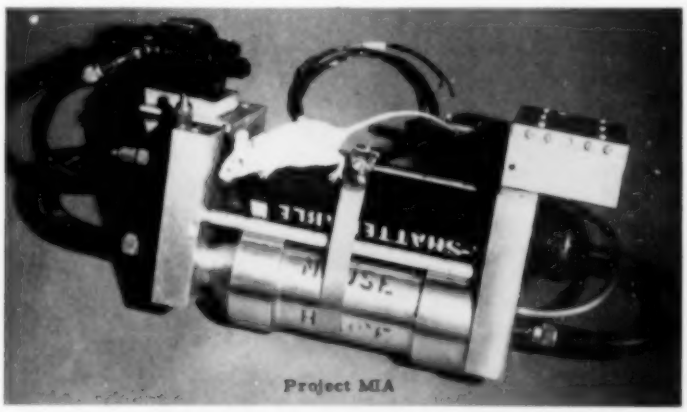


Fig. 1. Example of a package experiment.

As manned space systems become closer to reality, test vehicles having bioastronautics research as prime objective will be made available, as in the case of Project Mercury. It is the purpose of this paper to demonstrate the contributions which may be made by including bioastronautics package experiments as "secondary" or "hitchhike" experiments on vehicles now being tested, thus appreciably reducing the time and cost of development of advanced manned space systems.

In the past, there has been an understandable reluctance of missile engineers to become involved in biological research. Their objections have included the following:

(a) This experiment may interfere with or detract from our primary task of developing an effective weapon system.

(b) This is a job for the doctors, biologists and physiologists; let them do it.

(c) Animals are too hard to work with; they are unpredictable and temperamental and have no place in an engineering system.

In response to these objections, it may be pointed out that it is standard practice in missile development to test components of advanced systems as non-operational elements of the payload in current flight tests (e.g., open-loop tests of advanced guidance and propellant utilization subsystems have been incorporated in development flights of current ballistic weapon systems); that follow-on weapon systems to current ICBM's will very probably be manned,[2] and consequently operational life-support components must be made available in the near future; that the life-science specialists can develop, and are developing, these components but need engineering assistance to produce operational

units; that flight-testing of these components is an essential factor in their development which must not be neglected by the missile industry; and that, while animals are admittedly a new and unfamiliar factor in missile systems engineering, it would be difficult to find an unproven guidance system which is not more unpredictable and temperamental than a monkey . . . or a man.

An additional objection, and one strongly seconded by responsible workers in the bioastronautics field, is that the amount of information gained in any single flight test is seldom worth the effort associated with preparation for the experiment. Precisely because man, or any other animal, is a variable quantity, so that a single stimulus can produce widely differing responses from different subjects, and even from the same subject at different times, it is imperative that a large number of tests be performed to identify not only the statistical mean response but also the limiting extremes. The need for multiple testing also extends to the development of life-support equipment components. The load on the equipment will vary with the resultant effect of the numerous interconnected stresses acting on the subject at any given time, and the operational characteristics of the equipment are a function of the environmental regimes which the vehicle may traverse.

There are, then, two major advantages to the concept of package experiments for bioastronautics research:

(1) Package experiments are, by definition, minimum-interference experiments, and thus more acceptable to the program director whose prime concern is development of an effective weapon system.

(2) If properly designed, a package experiment may be flown on a variety of missiles with minimum modification, thus assuring a statistically significant data sample over a wide variety of vehicle-environmental conditions.

However, large numbers of tests are not enough. It is necessary that the objectives of any given program be clearly defined prior to flight testing of the package experiment, and that package-experiment results from various vehicle programs be collected at some central agency for correlation and evaluation. Furthermore, not only must there be consistency in evaluating and interpreting specific test results, but careful consideration must be given to the significance of the program objective as a contribution to well-integrated development of manned space flight systems.

II. AREAS OF BIOASTRONAUTICS RESEARCH

The term "bioastronautics research" includes the study of all factors which may influence the ability of man to survive and to function effectively in space. In the immediate future, the principal objective is solution of the problems affecting survival, and it is in this area that package experiments offer their greatest potential.

In general, bioastronautics research related to survival can be divided into three separate categories:

(1) Experiments to identify and evaluate the component factors of vehicle-induced and space environments. These factors include cosmic radiation, micrometeorite density and distribution, the thermal environment of space, and also the environment imposed during operation of the launching and re-entry vehicles, from the noise and vibration at liftoff to the deceleration loads at impact on return. These factors must be defined if suitable protection for man in space is to be developed.

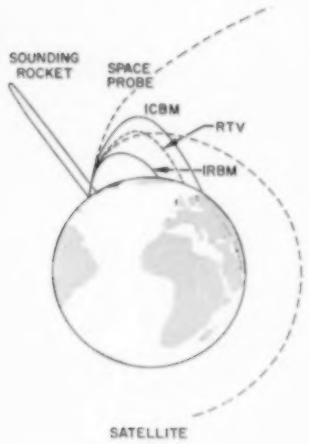
(2) Experiments to evaluate the performance of life-support component equipment under operational conditions. Subsystems such as air-purification units, temperature-control devices and the recovery and emergency-escape complex must be thoroughly checked out in the space environment before man is permitted to risk his life on their functioning.

(3) Experiments to obtain data to aid in formulation of predictions of man's

physiological response to the environment of space. As in other areas of medical research, much information on the responses of man may be obtained through experimental studies using lower-order animals. In general, the value of the data increases with the test-animal's biological similarity to man; thus monkeys are "better" than mice. However, there are a number of problem areas, such as the study of genetic effects, where small animals having short generation periods are most suitable.

III. VEHICLE CHARACTERISTICS

Since package experiments cannot be genuinely "non-interfering", it is important that the vehicle selected to carry the experiment be carefully chosen to produce optimum results in return for the effort expended. It is safe to say that all moderate- and high-performance rocket-powered vehicles can be used as carriers for bioastronautics experiments which will yield significant results. For example, there is a critical need for data on radiation encountered at altitudes traversed by most of our current development rockets. Also, during the ballistic



	TIME OF WEIGHT-LESSNESS	LAUNCH G LOADS	ALTITUDE	SPIN	RECOVERY	MAXIMUM VELOCITY
IRBM	~12 MIN	~10	400 NM	NO	(R & D VEHICLES) DATA CAPSULE	~15,000 FT/SEC
ICBM	~30-40 MIN	~7	800 NM	NO	(R & D VEHICLES) DATA CAPSULE	~22,000 FT/SEC
RE-ENTRY TEST VEHICLES	~10-40 MIN	~4-40	400 TO 1500 NM	NO	PROBABLY	15,000-24,000 FT/SEC
SOUNDING ROCKETS	30 SEC TO ~50 MIN	~4-40	UP TO 4000 NM (FAR SIDE)	NO	PROBABLY	—
SATELLITES	∞	~4-40	400 TO 22,000 NM	SOME-TIMES	PROBABLY	~25,000 FT/SEC
SPACE PROBES	∞	~8-40	∞	PROB-ABLY	NO	~35,000 FT/SEC

Fig. 2. Trajectory parameters of vehicles suitable for bioastronautics research.

At the present time, most of the experimental work related to man's survival in space has of necessity been conducted in simulated environments. These ground-laboratory experiments have supplied answers to many questions, and, even more important, they have served to define the problems yet to be solved. It must also be noted that many generally accepted "answers" are still unconfirmed, since the simulated space environment used in generating these solutions was based on speculation rather than definite knowledge. This aura of doubt and uncertainty must be dispelled before man is sent off into space; much of it can be removed by information obtained in a well-integrated program of package experimentation, with minimum expenditures of time and money.

portion of their trajectories, almost all rockets experience periods of weightlessness much longer than any which have yet been produced on Earth or in conventional aircraft flown along a parabolic arc. Experiments to investigate the effect of weightlessness on the operational characteristics of equipment developed on Earth, as well as the physiological responses of living subjects, can be performed on almost all rocket flights. However, on flights where spin stabilization is used, it is not possible to perform experiments which are dependent on weightlessness for effective results, since the vehicle's center of gravity follows a ballistic trajectory (and may be said to be "weightless") but an artificial gravitational field is created in all directions perpendicular to the spin axis by the centrifugal effect.

Similarly, there is a need for development of an effective recovery system and for definition of the environmental conditions associated with re-entry from a satellite orbit. Some advances can be made through experiments using relatively low-performance vehicles, but for really meaningful results the re-entry vehicle must enter the atmosphere at a velocity quite close to orbiting velocity, and at re-entry angles similar to those which will be used for manned vehicles.

A summary of the pertinent characteristic trajectory parameters of various classes of vehicles which may be used for bioastronautics research is presented in Fig. 2.

objectives and of the quality and importance of data which may be obtained. This experiment, originally proposed by Dr. John Lindner of the Space Technology Laboratories, Inc., has been flown on six vehicles to date, significant results being obtained from the two flights for which recovery was effected.

As shown in Fig. 3, two types of radiation packs were used; one is $1\frac{7}{8}$ in. $\times$ $1\frac{1}{2}$ in. $\times$ $\frac{3}{8}$ in. in size, weighing 1·4 oz, the other even smaller. The one on the left is identical with the pack incorporated in radiation badges worn by military personnel; this particular unit is shown sealed against moisture. This type of pack con-

Fig. 3. Radiation packs.

IV. SPECIFIC PACKAGE EXPERIMENTS

Let us now consider some specific examples of package experiments which, if incorporated in the development test flights of current missile systems, would provide experimental data of significant value in the achievement of manned space flight. Two packages in each of the three research categories are described.

A. Environmental Factors

1. *Radiation Package*

This experiment is an excellent example of minimum interference with primary mission

tains three radiation-sensitive films. Two of the films are sensitive to different levels of beta and gamma radiation, covering the range from 150 mr to 1000 r with an accuracy of 5–10 per cent. The third film is sensitive to neutron impact.

The pack shown on the right in Fig. 3 was developed at the Lawrence Radiation Laboratory at Livermore, California. The case encloses a stack of ten nuclear emulsion plates. The trace left by a charged particle in penetrating these plates is sufficient not only to permit identification of the cosmic particle but to allow

determination of its energy level as well.[3] From data recorded in the initial recovered sample of this experimental pack, the first positive identification of particles in the Van Allen radiation belt was made. Further valuable information is anticipated when the second recovered pack is fully analysed.

Since the nature and intensity of cosmic radiation vary not only as a function of time, but also as functions of latitude and altitude, the need for large numbers of flights to obtain fairly comprehensive data is evident. Fortunately, this can be accomplished quite readily. The film-badge packages cost only $1.50 each, or $1.00 per flight if recovered for re-use. They can be mounted easily, as shown in Fig. 4, without interfering with the location of other components of the vehicle.

Packages mounted as shown are suitable for use only on flights which go above 400 nautical miles altitude, into the regions of high-energy radiation (Fig. 5). It is only in such regions that impacts can be anticipated which exceed in energy the known attenuation caused by the shielding effect of the nose-cone shell. If a thin window were provided, perhaps on the nose-cone afterbody, valuable results could be obtained at altitudes as low as 150 nautical miles; cosmic ray surveys at these relatively low altitudes are of special interest since this is the region which will be traversed by early manned satellites.

Other improvements on the current package might include an indexing element, so that a sequence of packages could be exposed to the ambient radiation. Results could then be identified with specific areas of space, as determined by correlation of the trajectory elements with the time of exposure. By the present technique, the test results are obtained by comparison of the total radiation accumulated during the entire flight with the radiation recorded on an identical control package exposed on Earth. Additional refinements, which will lead to much more sophisticated results, can be incorporated with very small increases in the required weight and volume allowances. These improvements add complications, however, and are in no way a necessary condition for obtaining valuable results in future flights.

These radiation packs can be flown, with significant results, on all missions which go

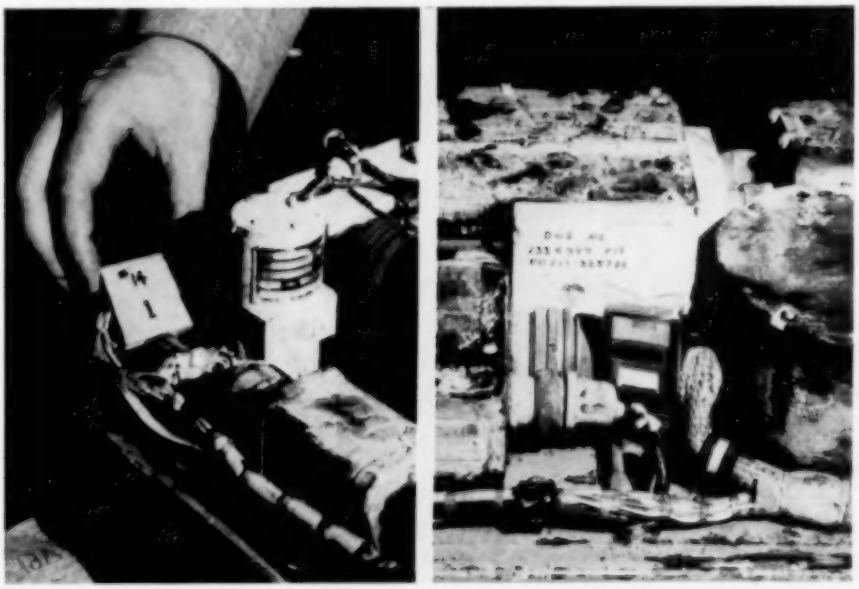


Fig. 4. Mounting of radiation packs in payload.

beyond the sensible atmosphere and which have a recoverable component (whether entire nose cone, data capsule or other element). The recovery equipment can also be eliminated if telemetry channels are made available, with moderate increases in weight and volume of the package experiment to accommodate the necessary instrumentation.

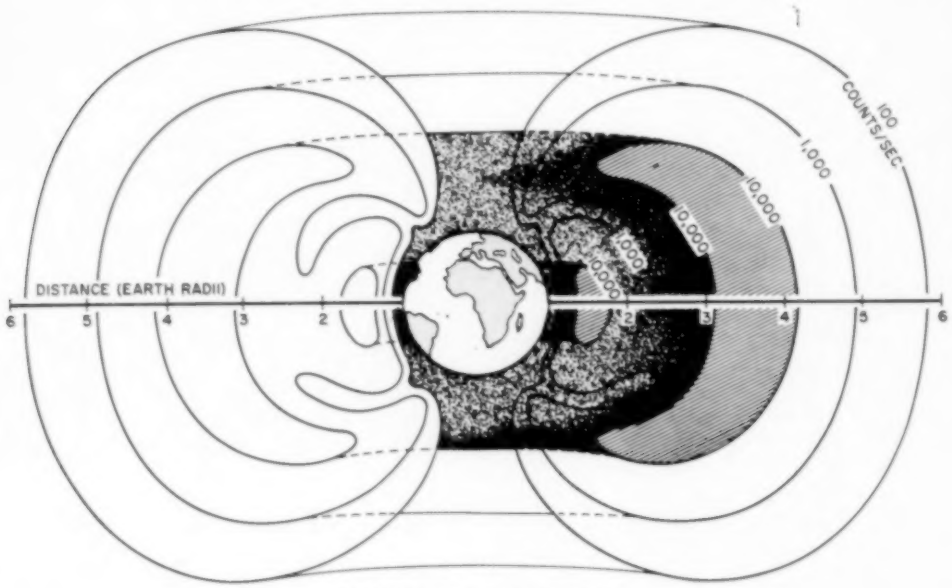


Fig. 5. Van Allen radiation belts.

2. *Vehicle Environment*

Many of the most severe stresses to which future astronauts will be exposed are associated with the conditions induced by the performance of the launching and re-entry vehicles rather than with the ambient conditions of space. The combined effects of noise, vibration, axial and lateral accelerations, in both the ascent and re-entry phases of the flight, may prove to be the critical stress condition of the entire mission.*

* It may be noted that temperature is not included as one of the critical vehicle-induced environmental stress elements. The state-of-the-art of heat shielding is such that no increase in temperature is anticipated within a nose cone as a result of external factors. The heat balance within a manned capsule is still a severe problem, but only as a function of the heat output of internal equipment components and of the man himself.

It is important, therefore, to identify and, where necessary, protect against these vehicle-induced environmental parameters.

It is common practice in missile engineering to investigate all of these stress elements in terms of their potential effect on hardware components of the vehicle systems. Unfortunately, human beings are seldom critically sensitive to stresses at the same levels as, say, a guidance system or a solid-propellant grain. Consequently, very little of the data currently available on the vehicle environment can be directly applied to the case where a man is included in the system.

The proposed package experiment to investigate these stress elements in regions of importance to a human passenger is shown schematically in Fig. 6. The sensor elements include axial, pitch and yaw accelerometers, a microphone to detect average noise level, and a vibration pickup for recording the low frequency vibrations (0–20 c/s) most critical for living organisms. By using a high sampling rate (20 samples/sec), sufficient acceleration and noise information can be obtained on a single commutated channel. Analysis of the vibration spectrum will require a continuous record of

that parameter. Thus, two channels of data are generated in the proposed instrument package.

The accumulated data can be returned for analysis in one of two ways. If the missile nose cone is to be recovered, a small recorder can be added to the package. The values shown for the recorder package correspond to the specifications of the Lockheed model M501 airborne magnetic recorder.

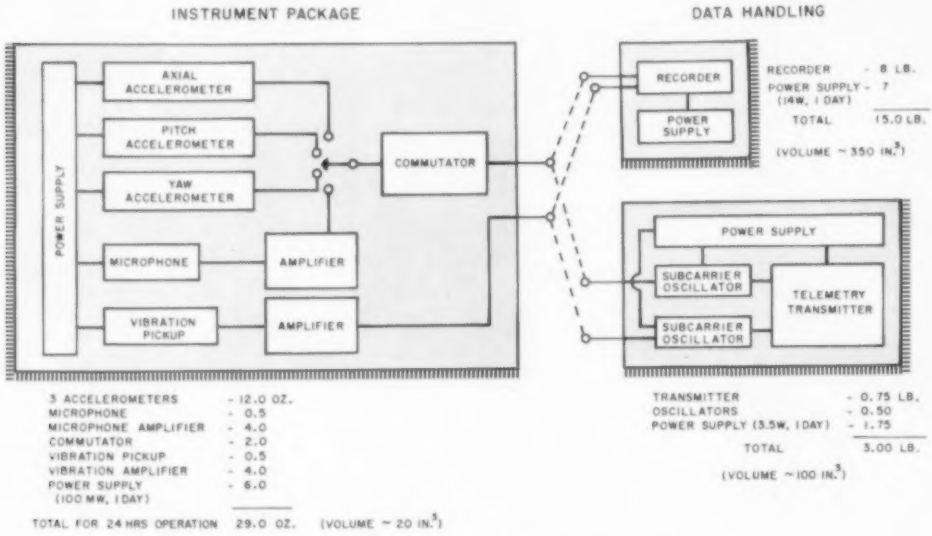


Fig. 6. Vehicle environment package experiment.

The second technique for obtaining the data is, of course, through telemetry, which must be used if the nose cone is not to be recovered. If sufficient channels are available on the vehicle's telemetry system, no additional weight need be charged to the basic instrument package for data handling. Alternatively, a separate two-channel telemetry transmitter, with power supply for 24 hr operation, may be added for about 3 lb and 100 in^3 additional allowances. One disadvantage of telemetry, as compared with recovery of a multi-channel recorder, is that data cannot be transmitted from the vehicle to ground receivers during a particularly interesting portion of the re-entry trajectory, when atmospheric friction causes ionization of the air-layer adjacent to the nose-cone surface.

The proposed instrument package can be flown to advantage on all missions using propulsion systems which may constitute stages of future manned systems, and on any mission having a re-entry trajectory similar to that which will be used on manned flights. For example, these data should be obtained for the Atlas and Titan in the immediate future and on any satellite or re-entry test vehicle having a relatively low-angle, low-deceleration re-entry trajectory.*

B. Equipment Components

1. *Escape Detection System*

An area of major concern in the development of manned space systems is the problem of assuring prompt separation of the manned compartment in the event of malfunction of the launching vehicle. A principal difficulty is to detect a failure (in any one of hundreds of thousands of component parts) which would require aborting the mission,

* Since the ascent and re-entry portions of the flight are of primary interest in evaluation of the vehicle-induced environment, it may be desirable to inactivate this package experiment during the orbiting phase of satellite flights in order to conserve recorder tape and/or power. This may be done by connecting the package to some suitable switching function in the vehicle system if it is not feasible to provide a separate signal element.

and to do this in time to eject the suitably protected space-man in a safe manner. By the nature of the problem, it is impractical—if not impossible—to test an emergency escape system by simulating all possible vehicle malfunctions. In order to reduce the required number of development tests, it is proposed that an escape-detection package experiment be devised which can be included in all test flights of all vehicle stages being considered for use in manned systems. The function and operation of this unit are illustrated schematically in Fig. 7.

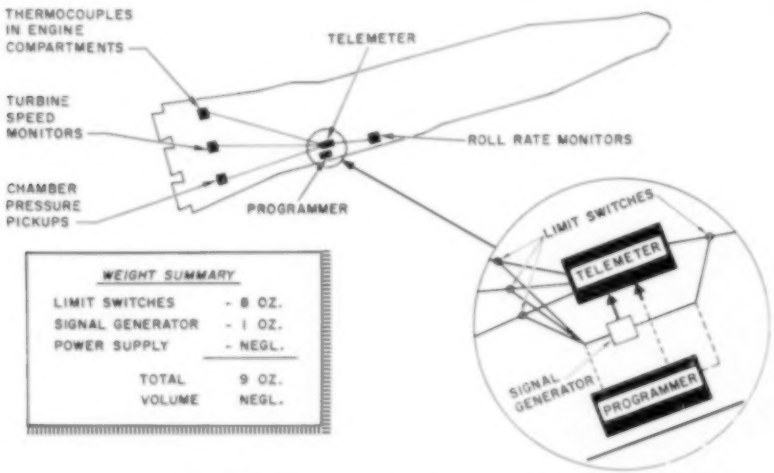


Fig. 7. Escape detection package experiment.

Most vehicle malfunctions fall into a limited number of classes. Specifically, imminent failures of the propulsion system are indicated by rough combustion, by turbine under- or over-speed, or by gross overheating in the engine area. Other vehicle breakdowns, such as structural failures or gross errors in electronic systems operation, are usually detectable in rapid roll rates about one of the vehicle axes, as picked up by the autopilot system. All of these potential-failure indications are normally measured and recorded throughout development test flights. Thus the proposed package experiment need consist only of limit switches placed in the circuits for measuring these few selected parameters, and a signal element (simulating the signal which would be sent to the programmer in a manned vehicle to trigger the emergency ejection system) to record on the telemetry "occurrence" channel the time at which the ejection would have been initiated. It would be desirable to have the emergency signal, generated by activation of one of the limit switches, go to the vehicle's programmer, which would then send an impulse to the telemeter. This impulse would be the same signal which would trigger ejection in a manned system, thus adding another element to the circuit being tested. Unfortunately, modification of a vehicle system programmer constitutes considerable interference, and, unless the programmer was quite similar to one planned for use in a specific manned system, the results would probably not be valid for this purpose. Consequently, routing the signal through the programmer is recommended only for specific vehicle flights, and not as an integral part of the proposed package experiment. In either case, the electrical signal to the telemeter is the only physical result of operation of the escape-detection unit; no actual separation or triggering of a firing squib is required. Consequently, the unit is very simple and light, and can be readily incorporated into current missile systems.

On successful flights, the unit will not operate; on partially successful flights, those for which

the malfunction would require emergency escape if the vehicle were manned, the ejection-simulating signal would appear on the telemetry record but the flight would not be compromised in any way by operation of the proposed unit. The information gained through use of the escape-detection system will supply answers to the following questions:

(1) Can all potentially destructive malfunctions be detected by measurement of a limited number of selected parameters? What are the allowable limits for these parameters, both to assure malfunction detection and to prevent unnecessary ejections?

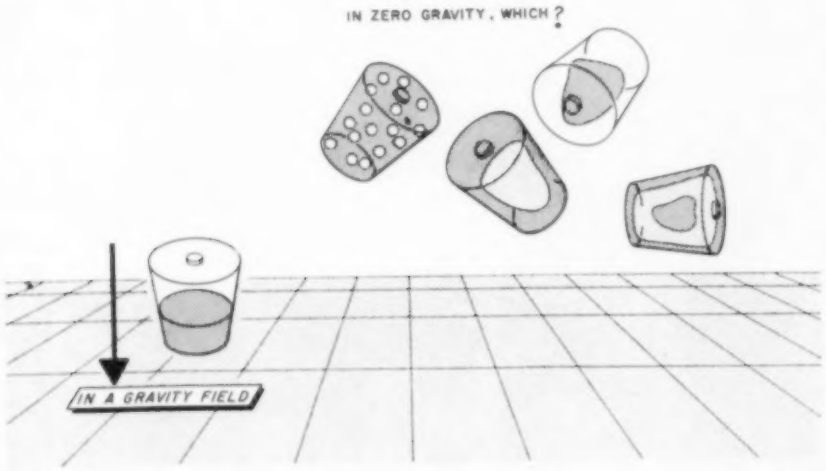


Fig. 8. Steady-state behavior of liquid–gas mixtures.

(2) What time increment is available between detection of the malfunction and complete destruction of the vehicle? Is it enough to effect escape of a human passenger?

(3) What are the extreme values of trajectory parameters which must be considered in designing an escape mechanism for each of the anticipated launching systems?

It is certainly an anomaly in missile development to propose an experiment which will give useful data only if the flight is a failure. In this case, however, such a negative approach is essential if safe recovery of the space traveler is to be assured under all conceivable—and some as yet inconceivable—emergency conditions.

2. *Liquid-Gas Systems*

Along with cosmic radiation and the element of hazard in space travel, weightlessness is one environmental condition which cannot satisfactorily be duplicated on the surface of Earth. Use of the parabolic-arc flight trajectory, with conventional aircraft, has enabled Air Force investigators to perform a number of highly significant experiments under zero-*g* conditions. However, maximum durations of weightlessness produced by this technique vary between 12 and 45 sec, so that these studies are necessarily limited to essentially transient effects.

One of the basic unknowns of weightlessness is the behavior of fluids; the nature of the problem is indicated in Fig. 8. There are quite valid arguments to support all these suggested possibilities for the behavior of liquid–gas mixtures; indeed any or all of these may occur under various environmental conditions and for various fluid properties. Unfortunately, the problem is not limited to methods of handling liquid nourishment which might be supplied to the astronaut. Liquid–gas interface problems are critical factors in the design of liquid-oxygen

converters (to supply breathable oxygen), water separators or condensers in temperature-control systems, waste-disposal systems and closed-cycle air-regeneration systems using algae, and—as an example related to vehicle performance rather than to the human payload—in liquid-propellant rockets on missions requiring re-start capability.

The obvious solution to problems associated with weightlessness is to create an artificial gravity field; the method most frequently proposed is to rotate the vehicle so that centrifugal forces would help keep things under control. For liquid–gas systems the density difference would cause the liquid to move to the outer edge of the rotating element, so that gas or liquid could then be bled off selectively from the center or from the periphery. However, for early, relatively small manned systems, rotation of the entire vehicle would have a severe disorienting effect on the human passenger—something like the effect of sitting on a spinning piano stool. The alternate method, rotating only the subsystems having liquid–gas interface problems, in a drum similar to an automatic washing machine, requires excessive power.

Before attempting to solve the problem, however, it would be advisable to define it more completely than has been possible in the past. For example, how is the static behavior of liquid–gas mixtures affected by such factors as viscosity and surface tension of the liquid, relative volumes, size, shape and composition of the container, temperature, and pressure? These questions can be answered by performing a series of simple experiments.

Consider the schematic drawing in Fig. 9. The transparent cube (about 3 in. per side is a reasonable size for an initial study) contains a mixture of a gas and a colored liquid. The container is sealed; its volume and the relative volumes of the contents are known. Thermocouples are mounted in the cube to give internal temperature variations. Three sides of the cube are backed with white non-reflecting paper. By use of suitably oriented prism complexes, images of the other three mutually perpendicular sides are aligned before a camera. These three sides are uniformally lighted and are marked with a grid to minimize distortion errors. The camera will record a triple image of the liquid–gas distribution pattern in three planes, at specified time intervals. For an initial test, these optical records might be made at 0·1-sec intervals

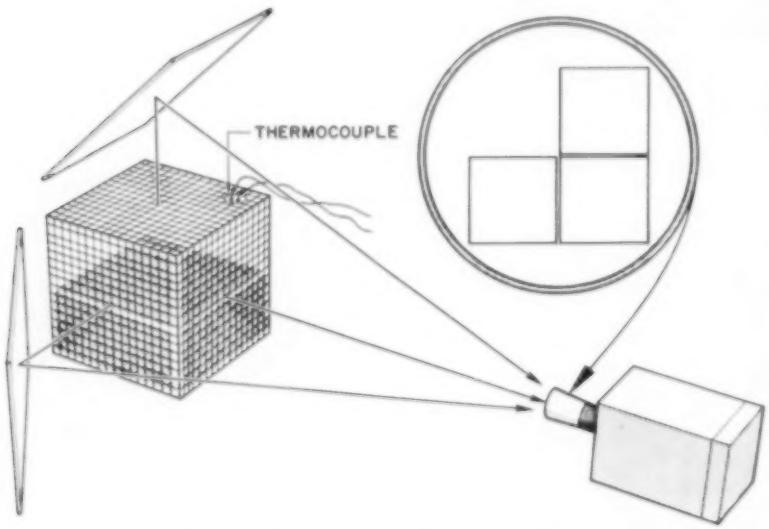


Fig. 9. Liquid–gas system package experiment.

during the period 30 sec before to 30 sec after burnout (beginning of weightlessness), and at 10 sec intervals thereafter.

The equipment necessary to perform the proposed experiment is largely dependent on the flight trajectory of the carrier vehicle. If the payload compartment is to be recovered, conventional camera and film may be used with no data transmission during flight. If the vehicle program does not include provision for recovery, the images can be televised for transmission to Earth. In the latter case, the volume and weight requirements are a sensitive function of the transmission distance, the resolution required, and the nature and number of receiving stations. The weight of the package experiment with television, including the transmitter and power supply, might be as high as 35–40 lb. This figure could be reduced if a vehicle telemetry channel were allotted for this purpose. If film is used, in recoverable vehicles, the net weight could be held to about 18 lb or less. In either case, the total volume requirement would not exceed 1 ft^3. These estimates of weight and volume requirements include allowances for all modifications necessary to test the potential effects on the liquid–gas system resulting from variation of the parameters mentioned above.

This package experiment can be flown on any vehicle whose trajectory includes a brief period of stabilized weightlessness. Since the zero-g requirement is a necessary condition for obtaining valid data, sounding rockets and short-range re-entry test vehicles are probably not suitable; satellite and space-probe missions would be ideal.

C. Physiological Research

1. *Discoverer Mark I*

The use of package experiments for physiological research is a currently accepted concept. In recent years, all known experimental data on biological responses to the environment of space have been obtained on flights of high-performance rockets whose primary mission was not to investigate bioastronautics problems. Project MIA (Mouse-In-Able) was a hitchhike experiment; the animal experiments in the Discoverer series were approved and developed on a secondary-objective basis; the dog, Laika, was also in all probability a "payload filler" in a satellite-performance test.*

Now that manned space flight seems a near-reality, is it worthwhile to continue flight-test experimentation with animals? It is the consensus among life-scientists that not only is it worthwhile, but—if space pioneers are to embark with any real assurance of safe return—it is imperative that physiological research using animals in space vehicles be continued and accelerated. Only through extensive use of space animals can we hope to protect the spaceman from potential malfunction of life-support equipment under load conditions, as well as from critical combinations of environmental stresses. Medical history gives ample evidence that results obtained through animal experimentation can be adapted for cases involving human subjects; the same is surely true for the problems of space flight.

Fortunately, the equipment and techniques needed for such experimentation are already well developed. The MIA package weighed less than 10 lb and occupied about $\frac{1}{4}$ ft^3 (with no allowance for transmitter or recorder, since the vehicle's telemetry system was used). The package was excessively heavy and bulky, mostly because of the need for using off-the-shelf items without modification in order to meet the 1-month development schedule. The same experiment could now be redesigned in a package about one-third the MIA package's weight and volume.

A more sophisticated experiment is the Mark I flown on Discoverer satellite vehicles, which carries four mice. The complete capsule, including the recovery system, weighs 195 lb and occupies almost 4 ft^3. However, the animal container and life-support components weigh only 15 lb and occupy about $\frac{1}{2}$ ft^3. Even these small allowances can be reduced. Considering the bioastronautics elements alone, this same experiment could be repackaged to fit in a space

* A possible exception is the series of sounding-rocket flights used by the Russians to develop techniques for recovery of their animal passengers.

of 600 in³ and would weigh only 12 lb, including power supply and amplifiers, but with telemetry transmitter or recorder requirements additional. The current Mark I test series could be extended by mounting this package in any vehicle whose flight program includes a weightless period of 15 min or more and, preferably but not necessarily, some provision for recovery of the mice. There is, of course, a limit to the amount of valuable physiological data which can be obtained from flight tests of "space mice". Even when this limit has been reached, however, modified versions of the Mark I capsule can still be used to provide a living load for development tests of improved equipment items.

Table 1. Weight Breakdown for Monkey Capsule (10 Days)

	Weight (lb)	Volume (in²)
Monkey and cage	8·0	600
Air purification system	10·5	300
Oxygen supply system	6·5	175
Temperature-control system	9·0	450
Waste disposal	1·0	450
Food (liquid)	6·0	450
Power supply (batteries, 5W)	17·5	300
Structure	3·0	negl.
Instrumentation	0·5	negl.
	62·0	2725 (1·58 ft³)

2. *Monkey Capsule*

The next step in space-animal research is to fly a monkey, since data obtained from primates are much more significant in terms of man than data obtained from mice. The Discoverer Mark II bioastronautics capsule is designed to carry a small primate instrumented for measurement of a wide variety of physiological parameters. At the time of writing, detailed information on weight and volume of the Mark II is not available in the unclassified literature, but some reasonable estimates may be made. Based primarily on the data given in Ref. (4), a summary of weight and volume requirements for a capsule containing one 6-lb monkey for a 10-day flight is given in Table 1. In comparison, the package experiment Bio-Flight Project No. 1 of ABMA, which was the first flight of a monkey in a Jupiter missile (and incidentally another hitchhike experiment), weighed 29·5 lb and occupied 670 in³ for a 24-hr operating life.

In addition to providing for physiological measurements, the proposed capsule can be modified to accommodate some very interesting behavioral experiments which have been developed by the Air Force School of Aviation Medicine. Furthermore, the proposed package experiment could be designed to accommodate various life-support equipment components, thus allowing flight-testing of components designed for use in manned systems under actual space-environment conditions and under load conditions similar to those which will be imposed by a human subject.

It would be of value to fly a monkey capsule on flights having weightless periods of 30 min or more, particularly on satellite flights where potential degeneration of physical condition or of performance, as a possible result of extended weightlessness, may be detected. It is highly desirable that primate tests include provisions for recovery.

V. CONCLUSIONS

It is common practice in missile development engineering to test advanced equipment components on-board flights of vehicles whose performance is not affected by their presence.

Results obtained from such minimum-interference, minimum-modification "package experiments" have contributed significantly to our progress toward space, both by increasing our basic knowledge and by expediting the development programs for advanced systems. Manned space systems are now under development, and the time has come to apply the "package experiment" technique to the field of bioastronautics research. Optimum results can be obtained only if a consistent, integrated experimental program is incorporated into the flight-test schedule of current vehicle systems, as opposed to the catch-as-catch-can approach which has yielded surprisingly significant but still inadequate results in the past.

To acquaint the missile engineer with the type of experimental data which would be useful, and also to give some indication of the degree of interference and modification which may be introduced by including package experiments in the test program, a number of typical experiments has been described. It has been shown that much valuable data can be obtained from these experiments, if they can be incorporated in flight-test programs currently underway. It is hoped that this opportunity to extend the scope of significant results of each and every space vehicle flight test will not be overlooked by those responsible for planning our advance into space.

REFERENCES

1. L. VAN DER WAL and W. D. YOUNG, *A Preliminary Experiment with Recoverable Biological Payloads in Ballistic Rockets: Project MIA*, Space Technology Laboratories, Inc. (Sept. 1958).
2. P. V. BARTLETT and R. A. FENLEY, *The Case for a Manned Space Weapon System*, Air University Quarterly Review, Vol. X, No. 4 (Winter 1958–59).
3. S. C. FREDEN and R. S. WHITE, *Phys. Rev. Letters* **3**, No. 9 (1959).
4. B. SIEGEL, *Survival Equipment for School of Aviation Medicine Experiment—Sub-System Selection and Weight Analysis*, Ramo-Wooldridge, Los Angeles, GM67.2-142 (July 1958).

A PARTIALLY CLOSED CYCLE LIFE SUPPORT SYSTEM FOR LONG-TERM SPACE FLIGHT

J. J. KONIKOFF
Missile and Space Vehicle Department, General Electric Company

Abstract—For man to survive in the hostile environment of outer space, he must be maintained in an environment that will essentially duplicate that existing on the Earth's surface. Studies have shown that with the exception of the manufacture of food, the problems connected with supplying all of man's necessities, i.e. temperature, pressure and humidity control, air purity, oxygen requirement and metabolic water supply, are amenable to solution using available knowledge. In particular the purification of air and the recovery of potable water and oxygen are examined in the light of their reclamation from man's biological wastes. A particular approach to the design of a partially closed cycle system is presented. Studies and experiments are currently in progress to solve some of the problems associated with this work. Results to date indicate the feasibility of the overall approach and further show that this particular system will probable be more efficient weight-wise than an open cycle storage system for space flights in excess of 15–30 man-days.

INTRODUCTION

The physiology of man demands that a certain minimum quantity and quality of water and nutrition be supplied him for his survival. When man is in his natural environment on the Earth's surface, the ecology existing there supports him with little effort on his part. However, should man attempt to leave the relative security of the Earth's surface, it becomes necessary to provide an environment that closely resembles that which he is leaving behind. Hence, it is mandatory that the daily requirement of food, oxygen and water be present as well as a means of disposing of his wastes. In this concept, man's wastes include such factors as exhaled breath containing water and carbon dioxide, urine, fecal matter and flatus containing toxic matter and perspiration.

Present technology limits the use of manned space vehicles to short duration flights in simulator, high-speed, high-altitude aircraft or balloon-supported gondolas. Technology in the fields of propulsion, guidance and structural design is progressing rapidly. In the near future, these technologies should reach a stage where it will be possible to place large payloads in space. An important facet of the overall technology required for manned flights of long duration is life support. Life support systems currently in use or presently being designed are based upon prestoring all of man's nutritional and respiratory needs and storing or discarding all wastes accumulated during flight.

AREAS OF INTEREST

The field of study encompassed with the life support studies includes two distinct disciplines, the medical and the physical. This paper is concerned only with the physical aspects of maintaining man in space. Of course, the physiological inputs required are obtained, as necessary, from the medical specialists. With the field of interest as now defined, Fig. 1 shows the various areas selected as being concerned with the support of human life in a hostile environment. All of the areas shown have been subject

1. Air Purification and Revitalization
2. Water Supply
3. Food Supply
4. Temperature and Pressure Maintenance and Control
5. Power Supply

Fig. 1. Life support studies.

to investigation for varying lengths of time. It might be generalized here that, with the exception of food supply, all are amenable to solution by application of known principles. In particular, the control of the temperature and pressure can probably be accomplished with available hardware or at worst some further development of this hardware. The power supply problem is being actively investigated based on the conversion of solar energy into a more useful form, i.e. application of the Seeberg and Peltier effects or thermionic converters, nuclear power sources, fuel cells, and extension of present day battery practice, to name just a few. The manufacture of food within a sealed cabin is a problem that to date defies solution. The use of algae or other biological sources has received much attention. Here, the concept simply stated, is to use man's wastes properly prepared to feed the algae and then to harvest the algae for food. In addition, since the algae converts CO_2 to oxygen during its growth process, a seemingly ideal cycle is formed. It has been estimated that 5 lb of algae per day will both feed a man as well as supply him with his oxygen requirement while man's wastes will provide sustenance to the algae. However, in order to cultivate this quantity of algae a large volume is required in addition to light energy in the visible wave lengths preferably at 6500 Å and at an intensity of about 500 f.c. Hence, it is obvious that additional work is required before this approach can be made to fix the weight, volume and energy requirements of a space vehicle.

In addition, research has been conducted in the field of hydroponics and artificial photosynthesis. Although hydroponic farming appears feasible and, in fact, may be developed to result in greater yields than conventional dirt farming, it is still difficult to apply this technique to open flight due to the water quantities required and the zero-"G" environment in space. Since the mechanism of photosynthesis is still not completely understood, it appears that a scientific breakthrough is required in order for man to produce artificially reasonable amounts of carbohydrate from CO_2 and H_2O. Thus, producing food in space is a problem that defies solution at the present time.

The air revitalization with oxygen and the water supply have also received considerable attention. However, to date, most of this work has been concerned with the supply of these materials for relatively short duration flights. However, as the exposure time in space becomes larger, it appears that the required weights of prestored food and water increases to a point where it becomes a major consideration. Thus, it becomes necessary to examine other possible means of supplying man with his needs for extended periods of time.

APPROACHES TO LONG TERM SPACE FLIGHT

Figure 2 lists three approaches to the solution of the oxygen and water supply for long term space flight.

1. Storage (Open Cycle) System Planned Depletion of Prestored Food, Oxygen, Water
2. Closed Cycle System Simulate Ecology Existing on Earth's Surface Including Food Production
3. Partially Closed Cycle System
   (a) Reprocess Man's Wastes to Supply O_2 and H_2O
   (b) Pre-Store Food

Fig. 2. Approaches to long-term space flight.

The first, the open cycle system, consists primarily of the prestorage of the required quantities of food, oxygen and water based on the premise of planned depletion as a function of time of flight. Using suitable desiccants or adsorbers, the contaminants introduced to the enclosed atmosphere may be controlled and by providing suitable containers, the urine and fecal matter may be stored until such time as they are discarded. Thus, for short flight time, the problem is relatively simple and may be solved by the application of an open cycle supply system.

Diametrically opposed to this is the second approach: the closed cycle system. Here the Earth's ecological system is simulated in such a

manner that oxygen and water are manufactured in some type of balanced system and man's wastes are used, possibly after processing, to support the growth of food. As mentioned earlier, man has not yet solved the food problem contained in this approach, hence, for now at least, this cycle must remain a goal to be reached for.

Obviously, in the wide spectrum of possible solutions, there exist many possible courses that will bring us somewhat closer to the ultimate goal expressed in the closed cycle approach. One such is No. 3, a partially closed cycle system, in which man's biological wastes are reprocessed to obtain his requirements for oxygen and water. Food, of course, is prestored.

MAN'S METABOLIC REQUIREMENTS

Before we can discuss ways and means for supplying man with his nutritional requirements, we must establish what these requirements are. If it is assumed that man requires a daily caloric intake of 3000 calories in food which will consist of 52 per cent carbohydrate, 23 per cent protein and 25 per cent fat by weight,[1] then using a somewhat simplified analysis, we may assume that this hypothetical food molecule has the formula $C_{15}H_{30}O_{10}N$. This molecule reacts in the process of human digestion with oxygen obtained from the atmosphere to form CO_2, H_2O and a hypothetical waste molecule CH_2ON as shown in equation (1).

$$C_{15}H_{30}O_{10}N + 16\tfrac{1}{2}\,O_2 \longrightarrow 14\,H_2O + 14\,CO_2 + CH_2ON \quad (1)$$

In this discussion the role of the trace elements, S, Ca, Hg, I, Fe, etc., will be ignored since their total weight will be small compared with that of any of the major constituents (C, H_2, O_2 and N_2). Based on the ultimate composition assumed for the food constituents, it can be shown that the food requirement amounts to $1\frac{1}{2}$ lb/day. Also, by simply balancing this equation, it is found that about 2 lb of O_2 are required for the oxidation process and the yields are very nearly 1 lb of water, 2·4 lb of CO_2 and about 0·1 lb of waste material.

It is of interest here to point out that man's metabolic process produces nearly 1 lb of water per day over and above the quantity of water taken orally.

Figure 3 presents an approximate material balance based on a 24 hr period. Examination of the input side shows the weight requirement for oxygen and water whereas the output side indicates a possible supply for these materials. They are, of course, the metabolic wastes—water and CO_2. Water, in this case, includes the water contained in exhaled air, perspiration, urine and the fecal matter.

Input (lb)			Output (lb)	
Water		5·0	Water	6·0
Oxygen		2·0	CO_2	2·4
Dry Food (3000 cal)		1·5	Solids	0·6
Carbohydrates	0·75		Other Gases	Negl.
Protein	0·35			
Fat	0·40			
Roughage		0·5		
Total		9·0	Total	9·0

Fig. 3. Approximate material balance (24-hr basis).
$C_{15}H_{30}O_{10}N + 16\frac{1}{2}\,O_2 \longrightarrow 14\,H_2O + 14\,CO_2 + CH_2ON$.

Also shown is a $\frac{1}{2}$ lb of roughage to be admixed with the food. This is included on the advice of the physiologists who have doubts concerning the possible effects on the gastro-intestinal tract as the result of little or no bowel movements, for an extended period of time.

POSSIBLE SOLUTIONS

The purpose of this section is to describe an approach to the solution of supplying man with his needs as regards to air purification, water and oxygen. Use is made of the procedure involving the reprocessing and/or regeneration of certain of man's wastes in order that a partially closed cycle can be achieved. Several possible solutions are available, however, early experimental results indicate that the approach described below appears to offer the highest potential of success due to its relative simplicity and ease of development.

A. Atmospheric Control

The removal of odorous material from the air

by use of activated charcoal appears quite feasible for prolonged space flight since mass requirements are small. Moisture control can be effected thru the use of a desiccant such as lithium chloride. The regeneration of LiCl requires a temperature of 300°F and a total of 1600 B.t.u./lb of water. In this manner, the atmospheric moisture is recovered. A molecular sieve appears suitable for the removal of carbon dioxide from the atmosphere. The regeneration process here requires a temperature of 200°F. In addition, 450 B.t.u./lb of CO_2 are required for desorption and 650 B.t.u./lb of CO_2 are required as sensible heat if it is assumed that the sieve is normally at 70°F.

B. Water

Man excretes water in his urine and fecal matter in addition to his perspiration and exhalation. Approximately 1600 g[2] are discharged daily in this manner. Although the recovery of the atmospheric moisture is relatively straightforward, the recovery of potable water from the urine and feces is a more difficult matter since these wastes contain toxic materials that are difficult to separate.[3] The approach related here consists of a process in which the raw material is initially vaporized. The vapor is then pyrolysed in the presence of oxygen and a catalyst. The resulting yield is then H_2O, CO_2, N_2 plus the oxides of the trace elements. Further purification of the water then may be accomplished, as necessary, either by redistillation, by ion-exchange or possibly by passing it over an activated charcoal bed. Initial experimental results, using urine only, indicate that the recovered water conforms to the U.S. public health service drinking water standards. Therefore, it appears that further treatment is unnecessary. Pyrolysis has been selected in this process since it is believed that the organic impurities will be converted to inorganic materials. The solid residue remaining in the still pot would eventually be jettisoned from the vehicle.

C. Oxygen

Earlier, it was pointed out that the human body in its respiratory process exhales H_2O and CO_2. Since both of these compounds contain O_2, it appears desirable to proceed in the direction that will result in the maximum separation of oxygen from these compounds. Hence, by electrolysis, the 1 lb of metabolic water can be made to give up all its oxygen, yielding 0·89 lb of oxygen.

The complete reduction of CO_2 to carbon and oxygen is a more difficult problem. However, partial reduction to CO may be accomplished with not too much difficulty. For example, using the H_2 obtained as a by-product from the electrolysis of water, the following reaction will result under the proper conditions:

$$CO_2 + H_2 \rightleftharpoons H_2O + CO \qquad (2)$$

$$2{\cdot}4 \text{ lb} + 0{\cdot}11 \text{ lb} = 0{\cdot}98 \text{ lb} + 1{\cdot}53 \text{ lb}$$

This is a modification of the water gas reaction and occurs at elevated temperatures (2750°F) over an iron chromite catalyst. This reaction must be cycled, approximately 60 per cent of the carbon dioxide being removed on this first pass. The water is then extracted and the cycle repeated until virtually all of the CO_2 is removed. Although carbon monoxide is generated by this reaction it need not necessarily produce a hazard since the reaction is carried out in a closed system which can ultimately be vented to outer space. Electrolysing the H_2O formed in the above reaction results in an additional 0·87 lb of oxygen for a total of

$$0{\cdot}89 + 0{\cdot}87 = 1{\cdot}76 \text{ lb approx.}$$

Since man requires about 2·0 lb of O_2 per day, by the method shown above, we have recovered over 85 per cent of his needs. The additional 15 per cent ($\frac{1}{4}$ lb) of O_2 may be obtained by the reduction of the carbon monoxide, although this reaction requires considerable investigation. A simpler means involves the reduction of hydrogen peroxide over a suitable catalyst. As shown below, about $\frac{1}{2}$ lb of H_2O_2 will yield the required $\frac{1}{4}$ lb of O_2 as follows:

$$2\,H_2O_2 \xrightarrow{\text{catalyst}} 2\,H_2O + O_2 \qquad (3)$$

$$0{\cdot}5 \text{ lb} = 0{\cdot}265 \text{ lb} + 0{\cdot}235 \text{ lb}$$

As a by-product, water is formed that can

serve any of several purposes, including a supply required for hygienic purposes. In addition, heat is generated which may prove useful.

Another method might be considered. As described previously, man's metabolic process an excess of water amounting to approximately 1 lb/day. Since 1 lb of water represents almost 0·9 lb of oxygen, after 1 day's flight sufficient oxygen can be produced to meet man's requirements merely by the collection of the additional water manufactured by man.

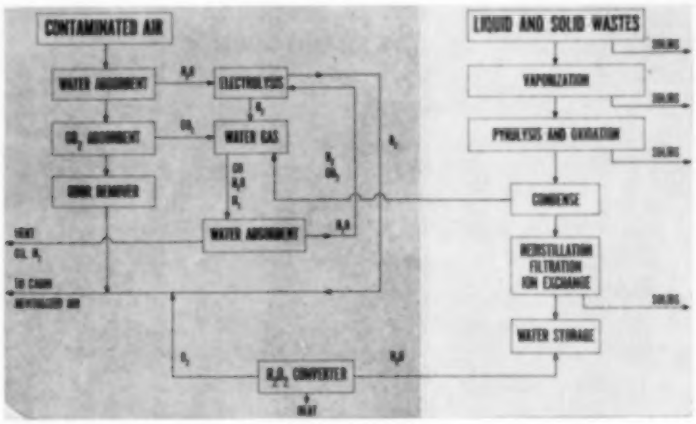


Fig. 4. Proposed partially closed cycle regenerative system.

Figure 4 is a schematic flow diagram illustrating the various processes described above. It shows the inter-dependence of the processes upon one another for their raw material. For example, following the flow lines concerned with air purification it can be seen that the water is adsorbed on LiCl beds, then released by heating whence it is electrolysed into O_2 and H_2. The H_2 is used to reduce the CO_2 obtained by regenerating the molecular sieve to obtain additional water. This water is then electrolysed, etc. This is intended to indicate that the entire reprocessing may be treated as one system composed of closely inter-related subsystems. The power requirements for this system are relatively high when compared to the open cycle system. Fig. 5 lists the energy requirements for the various subsystems. Also shown is the form of energy required, i.e. thermal or electrical. The total value shown, 30,000 B.t.u./day is possibly pessimistic inasmuch as no heat recovery is considered. It is possible that by proper design a fair quantity of heat may be recovered for further utilization, i.e. the water gas reaction is essentially thermoneutral; thus, as much as 10,000 B.t.u. may be recovered. Assuming that about 22,000 B.t.u./day are required (11,000 as thermal energy and 11,000 as electric energy), then, since the solar energy constant is 1·94 g cal cm^{-2} min^{-1} (10,320 B.t.u. ft^{-2} day^{-1})[4] and the following conversion efficiencies are attainable

Thermal	100 per cent
Electric	10 per cent

	Total Input (B.t.u./day)	Form of Energy
O_2	20,000	50% Electrical, 50% Thermal
H_2O	4,000	Thermal
Air Purification	6,000	Thermal
Total	30,000	
Solar Constant = 1·94 gcal (cm^{-2}) (min^{-1}) = 10,320 B.t.u. (ft^{-2}) (day^{-1})[17]		
1. Assume $\eta = 1{\cdot}0$ $A_S \approx 3{\cdot}0$ ft²		
2. Assume $\eta_E = 0{\cdot}1$ $A_{S_E} \approx 10{\cdot}0$ ft²		
$\eta_T = 1{\cdot}0$ $A_{S_T} \approx 2{\cdot}0$ ft²		
A Total ≈ 12·0 ft²		

Fig. 5. Approximate energy requirements.

the total surface area required to collect the solar energy becomes about 11 ft².

I would like to point out here that we are by no means advocating solar energy as the panacea to the power supply problem. Other sources may prove more desirable. The application of solar energy is included in this system primarily because it is "free" and available for long times.

ADVANTAGES

Obviously, a system as proposed above must be justified in order that it may replace the open cycle approach presently being used: What better justification could be found than that showing a definite weight advantage as a function of flight duration? To do this, calculations based on a number of assumptions were made in order that a comparison of the two systems could be made.

The assumptions for the open cycle system are:

(1) No regeneration is attempted.

(2) CO_2 and H_2O are adsorbed by LiOH and LiCl respectively. 2·4 lb of LiOH and 4·2 lb of LiCl are required per man-day.[5] Container weight is estimated to be 10 per cent of desiccant weight.

(3) All water is prestored. Thus 5 lb per man-day plus 10 per cent container weight are required.

(4) A liquid oxygen system supplied man's metabolic requirements of 2 lb of oxygen per day. The oxygen system is estimated to weigh 3·5 lb per man-day.

(5) Power requirements are estimated at $\frac{1}{2}$ kWh and weigh 7 lb.[6]

For the partially closed cycle system the assumptions are:

(1) The system as proposed is feasible.

(2) Since the desiccants used to adsorb the atmospheric moisture and carbon dioxide are regenerated periodically, only a little more than 1 day's supply of each is required. Adding a pound of charcoal for odor control, the total desiccant weight becomes 35 lb.

(3) All waste water will be recovered. Then, other than the first day's supply (5 lb) no storage is required.

(4) Oxygen will be recovered as indicated in Fig. 4. The make up required (about $\frac{1}{4}$ lb/day) will be obtained from the hydrogen peroxide decomposition. In this manner about $\frac{1}{3}$ lb of H_2O_2 per day is required plus the first day's supply of oxygen (2 lb).

(5) The desiccants will be reactivated by direct solar energy.

(6) Water will be electrolysed by energy obtained by solar batteries.

(7) At a power density of 13 lb/kWh[6] solar silicon cells will be used and the power package will weigh 130 lb.

Based on these assumptions, the following weight estimates have been obtained:

Open cycle system

(1) Air purification	8·5 lb/man-day
(2) Oxygen supply	3·5 lb/man-day
(3) Water supply	5·5 lb/man-day
(4) Pumps, tubing, power supply	15·0 lb
Total	15 lb + 17·5 lb/man-day

Partially closed cycle system

(1) Air purification and oxygen recovery hardware	66·0 lb
(2) Water recovery hardware	30·0 lb
(3) Power supply	130·0 lb
(4) First day's supply O_2 and H_2O + containers	7·0 lb
(5) Oxygen make up	$\frac{1}{3}$ lb/man-day
Total	233·0 lb + $\frac{1}{3}$ lb/man-day

The weight values tabulated above do not include food since, in either approach, food is prestored.

Figure 6 presents the results of this comparison. It can be seen that after a space vehicle flight time of about 15 man-days, it is justifiable to convert from an open cycle to a partially closed cycle life support system.

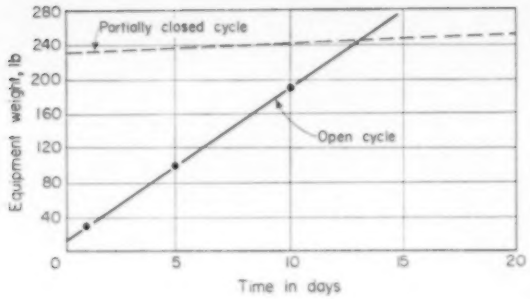


Fig. 6. Air–water open cycle–partially closed cycle break point.

RESEARCH AND DEVELOPMENT NEEDS

The processes discussed earlier are the result of computations and literature research. A considerable amount of experimental and developmental work is required in order to obtain verification of the feasibility of a recycling system.

Below are listed seven major areas requiring additional work.

(a) *Removal of impurities from air*

Although the techniques presently available appear to be adequate for this purpose, additional work is needed to optimize the system in order that a maximum quantity of impurities may be extracted from the air per pound of extracting agent. The major problem is to determine what impurities need to be removed and select suitable extractors for removing these impurities.

(b) *Conversion of carbon dioxide*

The water gas reaction has been studied extensively and is widely used in industry. A major area of interest here is the equipment design. It may be practical to remove additional oxygen from CO by reacting with hydrogen to produce water and carbon. The practicability of this application should receive further study.

(c) *Electrolysis of water*

Here again this process has been studied extensively and the design of the equipment is of major importance.

(d) *Recovery of water from urine and fecal matter*

An experimental program has been started in this area and, as described earlier, the results using urine alone indicate that potable water has been obtained without the need of further purification following direct condensation after pyrolysis. Presently, the investigation is concerned with recovery rates and energy requirements.

(e) *Oxygen from hydrogen peroxide*

Little difficulty is expected here since this process has been used extensively for many years.

(f) *Utilization of solar energy for operation of subsystems*

For certain types of operation where only heat is needed, the direct collection of solar energy by means of lenses or mirrors is quite efficient and needs little improvement. However, advances in obtaining electric power from solar energy could be substantial. Currently available solar batteries generally operate at less than 10 per cent efficiency and experimental studies indicate that perhaps no more than 20 per cent is available with present technology. Other methods of conversion do not currently offer comparable efficiencies. Research and development in the area of light weight highly efficient means of converting solar energy to electrical energy is needed. If a vehicle is oriented in space in such a manner as to direct one side continually toward the Sun, the opposite side will become quite cold due to radiation to outer space. Studies of the extent to which this type of

temperature differential can be utilized for power or be used for atmospheric purification might provide a solution to much of the problem of operating subsystems.

(*g*) *Food*

The supplying of food in any internal environment control system constitutes an enormous problem. The present as well as the advanced state of the art proposed here still prescribes storing the proper quantities of this material. However, a basic research area exists in the study of artificial photosynthesis. Equations (4) and (5) below

$$\text{Food} + O_2 \rightarrow H_2O + CO_2 + \text{Wastes} + E_1(\text{Energy}) \quad (4)$$

$$\text{Wastes} + H_2O + CO_2 + E_2(\text{Solar Energy}) \; \text{Food} + O_2 \quad (5)$$

represent, in its simplest form, the operation of the Earth's ecological system in which life is supported on a very large scale through the action of an approximately closed cycle involving a great complexity of biological material.

Equation (5) represents the photosynthetic process of growing plants while Equation (4) is the means by which animals carry out metabolic processes and as such is the reaction used by man to obtain energy for his activities. Although a man's energy requirements are about equal to the energy received by a square foot of surface normal to the Sun's rays at the Earth's distance, the efficiency with which he can collect and utilize this energy with the help of present technology is very low. Directly or indirectly he must utilize the second phase of the above-mentioned cycle. A field of corn, for example, utilizes on a year round basis, less than 0·10 per cent of the solar energy incident upon it. Recent studies[7, 8] using simple biological types (algae) have indicated the possibility of achieving an efficiency exceeding 10 per cent of the solar energy utilization.

One might solve much of the weight problem associated with prolonged space flight by utilizing a small short cycle closed ecological system in which all of the food requirements except perhaps a few of the more complex types (e.g. certain vitamins) are provided. The weight requirements of such a system would certainly have to be several orders of magnitude less than that of the system in which man is now nourished.

Basic research is currently being conducted along two paths aimed at solving this problem. One is the biological approach which utilizes small short cycle plant type organisms (algae, etc.) and has exhibited a higher utilization of solar energy than present food plants. In spite of these advantages, many problems must be solved before an ecological system based on these organisms is useable. Water requirements are excessive and processing equipment is much too complicated to be practical for space cabin application. The other approach involves accomplishing the same reaction without the use of biological material (artificial photosynthesis). Although this process has not been accomplished to date there is no reason to believe that it cannot be done, and research along these lines should be pursued more vigorously.

REFERENCES

1. S. Gottlieb, G.E. MSVD, Philadelphia, Pa., personal communication (Feb. 1958).
2. G. H. Bell, N. J. Davidson and H. Scarborough in *Textbook of Physiology and Biochemistry*, Williams and Williams (1950).
3. W. T. Ingram, *Orientation of Research Needs Associated with Environment of Closed Spaces*, AAS No. 57-17 (1957).
4. *Handbook of Chemistry and Physics* (39th Ed.), p. 3130, Chemical Rubber Publishing Co. (1957).
5. R. L. Schwaebel, Strato Lab. Development Mechanical Division of General Mills Co., Report No. 1648 (31 Dec. 1956).
6. J. C. Leisenring, *Considerations Concerning Power Sources for Satellites*, AIEE Paper No. 59-391 (Feb. 1959).
7. J. Myers, *J. Aviat. Med.* —, — (1954).
8. D. Burk, G. Hobby, T. Gaucher, *Closed Cycle Air Purification with Algae*, Paper presented at *International Symposium on Submarines and Space Medicine*, New London, Conn. (Sept. 1958).

SELECTION OF PHYSIOLOGIC MEASUREMENTS

FRANCIS BERCHMANS QUINN, JR. and VERNE O. BRECHNER

Department of Surgery, University of California, Medical Center

Throughout the clinical years of every physician's training one principle is virtually hammered into his habits of thought: the uncompromising necessity of direct personal observation of the patient. The young physician's eyes are trained to catch the subtle yellow taint of early jaundice, or the slight unsteadiness of the gait in disseminated central nervous system sclerosis. His nose is schooled in the appreciation of the acetone breath of incipient diabetic coma, the urea-like odor of kidney failure. Even before the stethoscope reaches the patient's chest, the physician hears the slight wheezing of bronchial obstruction, or the labored, bubbly breathing of cardiac failure. His touch tells him of fever, of muscle tone and of disordered thyroid function.

The manifold aids to diagnosis supplied by the biochemist and the electronics engineer have shorn many diseases of their mystery and elusiveness. Medical science has advanced beyond reasonable expectations as a result of these extensions of the doctor's eyes, ears and hands. Nevertheless, in the estimation of health, these devices remain always in the status of adjuncts to diagnosis.

Separate the physician from his patient by 100 miles of fast-vanishing atmosphere; tell him that the slightest decrement in his patient's physiological efficiency might result in the destruction of a multi-million dollar vehicle, and with it the patient himself; then charge the physician with the responsibility of detecting these changes or even anticipating them, with the full realization that upon his observations and judgment the man's life may depend. How then can he reach across the miles? What information can he gather which will permit him accurately to judge the pilot's condition? How much of the payload can he justly claim as essential to the requirements of physiological monitoring for the specific mission? Finally, how much weight and space should be allowed for the instruments whose telemetered signals would give him retrospective insight into the cause of failure should the vehicle be destroyed?

In the subsequent pages, we will investigate the techniques of physiological measurement which will apply to situations such as these, with a critical look at the value of the information thus obtained. In our personal view, there is good reason to suspect that the need for individual physiologic monitoring in space flight can be compared to the necessity of police surveillance in a society. As a social order evolves from infancy to maturity, the requirements for police activities generally diminish. As our space systems develop reliability and resourcefulness, the arguments of those who would have us strapped into multi-sensing harnesses, plugged into diagnostic computers and fed into therapeutic servomechanisms may lose some force. It may even become expedient, as an ultimate recourse, to combine these functions in a tightly looped, multiplexed, self-maintaining, miniaturized, feedback system employing molecular circuitry and existing in the somewhat exotic form of a living organism. In other words, we might have to send a doctor along.

Having committed ourselves to the task of obtaining information regarding the physiological condition of the man in the vehicle, we must first consider the people to whom this information will be of value. Stated simply, these people are three. The pilot is concerned, because this information relates directly to his chances

of collecting his pension. Next, the system builder must know if the man is functioning properly within the system and just how well he is functioning. The medical monitor wants to know his state of health. If the inhabitant is failing, he should be in a position to know what has happened, what can be done about it immediately, and how it can be prevented on the next trip. The eventual choice of instrumentation will evolve as an integration of these three points of view, with the pilot exercising veto power based on discomfort and interference with body movements, and the system builder, with his weight and balance slide rule, holding at bay the physician who secretly would like to measure everything from intraocular tension to anal sphincter tone.

What types of information can be obtained? Immediately we must distinguish between measurements of the environment; such as CO_2, humidity, etc., and observation of the physiologic responses of the individual himself; such as electrocardiographic configuration, pulse rate, ventilation rate, etc. Intuitively, one might conclude that as our manned systems achieve size, complexity and endurance, monitoring of the environment will supplant almost entirely the need for close, continuous medical observation of the individual crew member.

The first order of information is that which has instantaneous value; that relating, in descending order of refinement, to comfort, functional capacity, consciousness and life. The second order of information could be termed prognostic, that which permits detection and evaluation of conditions presenting as threats to continued successful functioning and existence of the human component. The third order of information can be called retrospective; that which, in the event of partial or total failure of the system, gives opportunity to identify correctly the probable causes of such failure.

How shall we get this information? That relating to comfort, function, consciousness and life is quite simple to obtain. Two-way voice communication together with a telemetered electrocardiographic signal will tell us what we need to know. The individual can advise us if he is hot, hungry, nauseated, dizzy, tired or in pain. From his conversation we can judge whether he is calm, rational and perceptive, or irritated, impatient, obtunded, in panic or delirium. Should voice transmission cease, the electrocardiograph will signal whether the organism continues to exist, and will give some idea of its salvage value.

Prognostic information requires more complicated data. It is secured by instrumentation of both man and his environment. The choice of instrumentation is best based upon a review of the altered conditions which the man can be expected to encounter, such as hypoxia, dysbarism, fatigue, heat stress, toxic substances, accumulations of metabolic products, acceleration or lack thereof, emotional stresses, noise and vibration, and ionizing radiation effects. The values obtained through this sort of instrumentation can be compared to established time–intensity tolerance limits for man, and a reasonable expectancy for continued function and perhaps recoverability can be achieved. It is conceivable that certain alterations in the mission profile might be dictated by this sort of information.

The final order of information, that with chiefly retrospective value, is not so easy to define. As in solving a crime or studying an aircraft accident, every scrap of information has the potential of being the one clue which will unlock the mystery. Instrumentation in this category represents somewhat of a scientific luxury, in that the data obtained would relate to adverse conditions over which neither the occupant nor the Earth-based controller exercised any significant degree of remedial control. Examples of this would-be data regarding intestinal motility or labyrinthine responses during prolonged weightlessness or tumbling, or a device to signal the occurrence of sudden, catastrophic capsule decompression or excessive ionizing radiation.

As a final word of explanation in reference to these categories of information, let me say that they apply not to instrumentation devices or systems as such, but rather to the intelligence obtained thereby. As you have doubtless

observed, there is much overlapping among these classes of information, but we believe that this method gives one a push in the direction of honest, clear-headed appraisal of the relative value of individual measurements.

What techniques have we for appraising the occupant's condition and the suitability of his environment? You have probably already become aware that a physician is never at a loss when required to prescribe diagnostic tests and laboratory determinations. From the pulse rate to the ballistocardiogram, he has at his disposal an almost infinite spectrum of physical, biochemical electrophysiologic measurements from which he can draw information of varying significance concerning the condition of the human organism under study. The next portion of the discussion will pertain to the measurements which seem likely to be employed in space flight, based upon their values to the entire system, and considered in the light of the significance of the data, the physical characteristics measuring equipment and their particular disadvantages, such as telemetry requirements or discomfort to the subject.

In a generally descending order of utility, we have electrocardiography, two-way radio communication, television, ambient gas pressure and composition, body surface and interior temperature, oximetry, peripheral blood pressure, and ventilometry and electroencephalography. Shading off into the region of scientific luxury, one goes on to list galvanic skin resistance, continuous tissue metabolic rate, pupillary size, eye-blink frequency, electro-oculography and gut motility. Performance tests such as simple arithmetic problems, tracking tasks, coordination and muscle strength tests, as well as measures of various auditory and visual perceptive thresholds fall into this latter category, although some of them could be included at no weight penalty to the system.

TWO-WAY RADIOTELEPHONE

Man is unique among experimental subjects as well as control systems in that he can be trained to observe and report enteroceptive and exteroceptive sensations. In this regard, the monitoring flight surgeon is at a distinct advantage over even his clinical counterpart, the anesthesiologist, who, although he sits next to his sleeping patient with his fingers on the pulse, must depend entirely upon objective signs with which to assess his subject's condition. In the Litton inhabited vacuum chamber, the intercommunication system demonstrated itself to be the most sensitive indicator of fatigue, anxiety, irritability, judgment deterioration, dysbarism and various types of discomfort. With voice communication requiring a bandwidth of about 3000 c/s, continuous verbal reporting by the occupant of a space vehicle is certainly possible in cis-lunar space. Full utilization of this channel, from the standpoint of total system monitoring, as well as physiologic surveillance, suggests that possibly our pilot might be trained continuously to verbalize his ideation, even when intensely engaged in his tasks. In other words, after the fashion of a boxing announcer, he could keep up a continuous chatter, relating observations pertaining to his internal and external body perceptions, his environmental control system, as well as his assigned tasks in astronavigation and scientific measurements. The ground observer, already personally familiar with the occupant and with his particular verbal mannerisms, through experience in the simulation laboratory, would listen, with an ear tuned not only to content of the message, but also to more subtle features, such as inappropriate humor, verbal blocking, perseveration, bizarre thought content and flight of ideas. With so much attention being focused on the development of devices which will tell the pilot that he is getting sleepy, becoming nauseated, developing a headache or becoming too warm, it is well to remember that sensing, reporting and corrective systems have already been included in this particular component, and at no extra charge.

ELECTROCARDIOGRAPHY

Of all instrumentation applied to humans and animals during high altitude flight, the electrocardiogram has enjoyed the most universal acceptance. The electrocardiograph has also been extensively applied in the operating room

as a continuous monitoring system during surgical operations. In this latter situation the changes in heart muscle potential accompanying cardiac systole may be amplified by transistors and fed into the circuit of an ammeter and earphone. Thus, each systole is registered as an audible click and a deflection of the meter indicator. The observer is immediately aware of sudden changes in cardiac rate or rhythm. Such a cardiac monitor is light in weight, rugged in construction and powered for long periods of time from a single 1·2 V dry cell.

A more sophisticated method of cardiac monitoring is the visual recording of the amplified potential alteration accompanying heart muscle contraction. Continuous visual recording may be accomplished with a cathode ray oscilloscope or a pen-writing galvanometer. Alterations in the electrocardiogram may indicate toxic states and change in heart position, as well as disturbances of cardiac rate and rhythm. Inasmuch as the function of the hypoxic or arhythmic heart is usually impaired, the electrocardiographic signal may thus reflect impaired function. This signal, however, is still simply an alteration of electrical potential during myocardial contraction and relaxation. It is not a measurement of the strength of contraction nor a measurement of the amount of blood propelled along the circulatory system. Indeed, in extreme circumstances in both the operating room and laboratory, it is possible to view a heart which rhythmically signals a fairly normal ventricular contraction complex but does not visibly contract.

GAS MONITORS

The gas components within the suit atmosphere may be monitored by physical methods which are employed in the surgical operating room.

Carbon dioxide concentration in the patient's inspired and expired gas may be monitored during surgery by measuring the amount of infrared radiation absorbed by the gas mixture. The variations in CO_2 concentrations as the patient breathes may also be used indirectly to observe respiratory rate and to determine that gas is actually moving in and out of the patient's lungs (i.e. CO_2 concentrate in ambient air is about 0·04 per cent while the concentrate in gases expired from the lungs is between 4 and 5 per cent). An index of the adequacy of pulmonary ventilation is achieved by observing variations in the peak expiratory CO_2. Respirations that are either too slow or too shallow increase the peak expiratory CO_2 while respirations that are too rapid and deep decrease the CO_2 concentration. The infrared CO_2 analyser may be applied to a pilot by insertion of a small soft catheter to the nares or mouth, or by incorporation into a helmet or mask. Oxygen is usually measured in the operating room by devices which sense the paramagnetic quality of the gas mixture. A gas sample is conducted into a chamber containing a test body suspended by a quartz fibre in a magnet. Such a device possibly would not remain operational if subjected to rapid acceleration and deceleration of launchings and landings.

Oxygen absorbs energy in the ultraviolet range and could be analysed instantaneously by a device similar to the infrared analyser but employing ultra-violet absorption.

Thus far we have assumed that oxygen and carbon dioxide are the only gases of interest. It seems possible that other gases, perhaps toxic in nature, may appear in the suit atmosphere and that a monitor capable of detecting the presence and concentration of several such components simultaneously would be desirable. Recent advances in the field of gas chromatography may provide just such an instrument.

TEMPERATURE

Internal body temperature remains constant under a wide range of environmental temperatures. Departures from normal in an otherwise healthy individual indicate failure of the homeostatic mechanisms concerned with maintaining heat balance, and generally connote a brief expectancy for the organism. Body interior temperatures compatible with life range from 4° to 42°C, while temperatures consistent with satisfactory human performance lie between 34° and 40°C.

Skin temperature represents the combined effects of environmental temperature, skin evaporation and peripheral blood flow and, as such, seems to be a somewhat ambiguous reflection of physiological status.

OXIMETRY

Oximetry refers to photoelectric measurement of the transmission of red light at a wavelength of 6400 Å through a relatively translucent vascular appendage such as an ear. This technique gives instantaneous and continuous monitoring of the blood oxygen saturation. Oximetric analysis depends upon the fact that oxyhemoglobin transmits 70 per cent of incident red light of 6400 Å, while the transmission of light at this wavelength by reduced hemoglobin is negligible. The ratio of oxygenated to reduced hemoglobin is thus one of the factors influencing the transmission of red light through a vascular bed, while the other variable is the amount of blood contained within the vascular structure. According to Beer's Law, the absorption of light is directly proportional to the amount of material containing the absorbent that must be penetrated by the light, as well as the concentration of absorbent in this material. Thus, the total amount of blood in the sample as well as the amount of reduced hemoglobin must be taken into account. Fortunately, near-infrared light, of 8000 Å is absorbed by oxyhemoglobin and reduced hemoglobin alike. Therefore, it is possible to correct for the total amount of hemoglobin present, using proper optical filters and electronic circuitry. Two photoelectric cells, one with a filtered peak sensitivity at 6400 Å and the other with a peak at 8000 Å, are employed to accomplish the necessary correction. The electrical current resulting from the illumination of the 6400 Å photocell through the vascular bed, is balanced against the current generated by similar illumination of the 8000 Å photocell. The resulting differential current is amplified to drive a pen-writing galvanometer. The deflections of the pen are relatively unaffected by the changes in total hemoglobin content, and hence relative changes in oxygen saturation of the blood stream are accurately recorded.

This type of instrument has considerable usefulness in situations in which hypoxia might be expected, such as in anesthesia or in the testing of oxygen-breathing equipment. The sensing unit is generally worn on the ear, and while not particularly uncomfortable for short periods of time, it is easily disturbed and can become quite annoying in protracted applications.

BLOOD PRESSURE AND PULSE

The importance of monitoring peripheral arterial blood pressure variations is readily recognized both in the operating room and in space programs. The simplest clinical method of measuring blood pressure involves the application of an inflatable cuff to the proximal portion of an extremity. The cuff is inflated and a stethoscope is placed over an artery distal to the cuff. As the cuff deflates an audible pulse appears when the cuff pressure falls below peak arterial pressure. Although blood pressure may be monitored directly with a strain gauge recording pressure from an endarterial catheter, the practical method is similar to the classical method first described. A proximally placed inflatable cuff is employed as before, and the stethoscope is replaced by a magnetic or carbon microphone taped firmly to a finger. With the cuff deflated, each heart beat causes a slight pulsatile expansion of the finger which is detected by the microphone. When properly amplified and fed into an ammeter or earphone, the pulse is registered as a detectable click, or deflection of the ammeter, similar to the cardiac monitor previously described. It should be noted that the significance of this signal is different from that of the signal from a monitor driven by changes in myocardial muscle potential. The pulsatile expansion of the finger is evidence that heart contractions have propelled a column of blood through the circulatory system. At specified time intervals the inflatable cuff is filled to a pressure adequate to compress the major arteries and thereby obliterate the finger pulse. Systolic blood pressure is recorded as the maximal cuff pressure necessary for disappearance of the finger pulse. During the interval

between blood pressure measurements, the pulse monitor is used to observe pulse rate and rhythm. A refinement of this method, although not employed in the operating room, is the replacement of the arm cuff with a small variable pressure cuff on the monitored finger. The transduced signals from finger pulsations are fed into a servomechanism which adjusts the finger cuff to systolic pressure. Obviously, the servopressure regulation should not be used continuously as it would effectively obliterate blood supply to the finger.

RESPIRATORY MONITORS

The usual method of monitoring respirations during previous space probes has been by means of a strain gauge attached to an elastic chest band. This arrangement records the chest motion associated with breathing. Thus the respiratory rate and a crude estimation of the vigor of each respiratory motion is recorded. It should be realized, however, that movement of the chest cage does not insure the exchange of gas in and out of the pilot's lungs. Anesthesiologists are well aware that a patient's airway may become obstructed when relaxation accompanying unconsciousness allows the tongue to fall back into the pharynx. In such a situation, vigorous respiratory movements will ensue for several minutes, although no air enters the lungs. A pilot, therefore could die of respiratory obstruction from any situation causing a degree of unconsciousness deeper than normal sleep even though the condition responsible for the loss of consciousness were not in itself lethal. Differentiating such a death of respiratory obstruction from death caused by a situation which is lethal would be of importance in reconstructing the true mechanism of a space tragedy. If during early space probes an astronaut dies and the mechanism of death cannot be readily explained, there will be a great temptation to invent a new mythical syndrome, such as "space shock". It is interesting to note that anesthetists labored under the mythical syndrome of "*status thymolympthaticus*" for years. This diagnosis was employed to explain unexpected death on the operating table. It is now realized that deaths alleged to *status thymolymphthaticus* were in all likelihood due to the simple mechanism of the relaxed tongue falling back into the pharynx and thus choking the patient.

A monitor which employs the movement of gas in and out of the patient's airway would provide more information than the elastic chest band. Such a monitor would not only record respiratory rate, but also the tidal and minute volumes. Two gas flow monitors frequently used in the operating room which might be applied to a pilot's airway are the venturi flow volumeter and the screen orifice flowmeter. Both of these meters operate by calibrating the drop in pressure as gas passes through the meter with volume flow. Such meters, however, incur the disadvantage of increasing the resistance against which a pilot must breathe. Although the resistance is unnoticed for short intervals, it could prove fatiguing over prolonged periods of time.

ELECTROENCEPHALOGRAPHY

The electroencephalogram records the minute potentials emanating from the cerebral cortex. Alterations in the frequency and voltage of these potentials are sensitive indications of the degree of electrocortical depression. In general, as cortical depression progresses from the activity characteristic of wakefulness through drowsiness of fatigue and light physiological sleep to increasingly profound degrees of unconsciousness, the frequencies of the brain waves slow and the voltage of each wave form at first increases and, as the cortex becomes severely depressed, finally decreases until no detectable electrical activity is recorded. The various patterns associated with different degrees of cortical depression are characteristic enough that an anesthesiologist can diagnose from them the planes of surgical anesthesia. Cortical depression may result from anesthetic agents and other noxious gases, hypoxia, decreased circulation of blood to the brain, hypothermia or increased cerebral venous pressure. It should be noted, however, that although the patterns associated with different degrees of cortical depression are characteristic, there are no

distinguishing features which differentiate one depressing agent from another. Thus, if an anesthesiologist observes markedly depressed electrocortical activity, he must quickly decide whether the electroencephalographic changes are caused by anesthetic overdose or one of the other cortical insults enumerated. In a similar situation, an observer viewing the electroencephalogram of a subject in a space capsule might detect and warn the subject in the early onset of light electrocortical depression but be unable to determine from the electroencephalographic patterns alone precisely which of several possible factors was responsible. Although information from other instrumentation might aid in identification of the offending depressant (such as fatigue or hypoxia), it is quite possible that the most valuable diagnostic information would be gained by interrogation of the subject. Since judgment and consciousness are impaired with relatively slight cortical depression, the duration of time between onset of electroencephalographic change and deterioration of observer–pilot communication may be brief. The significance of this critical time lapse is appreciated all the more when it is realized that correction of the factors responsible for the change might depend upon the pilot's ability to respond to the observer's instructions. Perhaps the most valuable use of the electroencephalograph would simply be in differentiating between unconsciousness of light "physiological" sleep from which the pilot can be aroused, and the more severe pathological depression of a true brain insult.

GALVANIC SKIN RESISTANCE

Galvanic skin resistance determination refers to the measurement of point-to-point resistance on the surface of the body. It is not known exactly what produces the changes in skin resistance, nor has the specific significance of this quantity been elucidated. Nevertheless, we believe that the galvanic skin resistance is controlled by the reticular formation in the central nervous system, and indicates the general degree of arousal of the subject. Actually, two measurements are made: the spontaneous or nonspecific skin resistance which changes as a result of internal or undetected stimuli, and resistance changes resulting from specific stimuli of specific intensity, usually in the form of an electric shock. It has been shown that as the state of arousal or general alertness increases from drowsiness to attention to extreme excitement or panic, the occurrence of nonspecific resistance changes increases in frequency. The responses to specific stimuli increase in amplitude up to the attentive state, then decrease in amplitude as excitement heightens. Environmental temperature can alter the skin resistance greatly, and there is considerable variation in measurements among individuals. Nevertheless, a fairly discrete pattern evolves for the individual, and it has been suggested that such a device might be expanded into an automatic sleep-warning system, in the hope of preventing in space travel the kind of catastrophes which occur on our highways when drivers fall asleep at the wheel. Weight and power requirements for such instrumentation may be quite small, in that skin resistance measurements can be combined with the electrocardiographic monitor.

CONCLUSION

In conclusion, we would like to suggest that the selection of monitors might rationally progress in the same order as a physician chooses his methods of patient observation.

First on the list would be direct audio contact with the pilot as provided by two-way radio. By these means the pilot's mental attitudes, judgment and response to various situations could be directly observed. Such factors as fatigue and mild hypoxia or toxic response to noxious gases would possibly first be noticed by observing slurred and slowed speech or abnormal behavior. Indeed, hypoxia has been diagnosed in high-altitude chambers by the subject merely mentioning that the chamber lights seemed to be dim.

Electrocardiographic monitoring is essential because of the very fundamental significance of the signal as a life indicator, as well as the simplicity and reliability of the technique. The television cockpit scanner will yield information

of direct value to other scientific disciplines as well as to the physician. The advantage of having a close view of what is happening in space needs no clarification. Measurements of the environmental conditions such as cabin pressure, temperature, humidity and gas composition, are in a sense indications of the health of the life-supporting system. While these are essential beyond question, the necessity for telemetering all these measurements is not clear, especially when the human occupant could monitor these just as well.

A final point should be offered in support of the use of standard clinical monitoring devices such as blood pressure level, ventilometry, electroencephalography and oximetry. The various physiological catastrophes to which man is subject manifest themselves in well-documented patterns of response. Although there may appear to be some overlap in the applications of these several measurements, the relative point of onset and rate of progression of disturbances in these basic functions could offer much in the retrospective analysis of unsuccessful ventures.

IMPACT TOLERANCE OF MAMMALS

M. KORNHAUSER and R. W. LAWTON
Missile and Space Vehicle Department, General Electric Company

Abstract—Correlation of experimental data on human impact tolerance, treating the test subject as a typical structural element, has led to the conclusion that impacts at velocities below about 80 ft/sec are tolerable regardless of the peak deceleration. The present paper is concerned with the empirical verification of the theory adopted in this study of human acceleration resistance, using mice as test specimens.

Impact tests were performed using a drop tester capable of velocity changes up to about 90 ft/sec, with various stopping devices for shaping the acceleration-time pulses. The practical ranges of impact parameters for the mouse "impact sensitivity curve" were covered by varying the impact duration from less than 0·5 msec to about 2 msec and peak accelerations from about 500 *g* to about 10,000 *g*. The effects of acceleration, impact duration and velocity change were evaluated.

The conclusions reached in this mouse impact study, when extended to larger mammals and humans by thorough experimentation, will have immediate application in the establishment of a simple design criterion for manned satellites and space vehicles on landing impact.

INTRODUCTION

Within the past several years a considerable body of experimental data has been generated on man's tolerance to acceleration or deceleration. The largest portion of this experimentation has been performed by centrifuge; essentially a steady-state loading. Some work has been done with fairly short-duration centrifuge spins and, more significantly, with shorter duration loadings approaching the impact situation through use of sled tracks and other impact test devices. Although it is recognized that the duration of acceleration is an impact parameter which ranks in importance with peak acceleration, a clear understanding has not been developed of the relative roles of peak acceleration, shape of acceleration-time pulse (related to onset-of-acceleration) and duration of acceleration.

In order to permit intelligent designs of vehicles such as manned satellites, an attempt[1] has been made to extend our understanding of man's acceleration resistance to the regions of short-duration impact. The approach was to borrow a structural loading-response model valid for inanimate structures, and to correlate the available experimental data on human impact resistance within the framework of the structural model. Although the human impact behavior appeared to conform to the model of structural behavior, there were far too few datum points to permit the use of the results with any reasonable level of confidence. It was concluded, therefore, that experimental data on acceleration tolerance of man in the short-duration impact region must be obtained on a firm statistical basis.

The one feature of the structural-human model which is of greatest practical significance is the prediction of the existence of a minimum impact velocity (for humans, about 80 ft/sec) below which there will be no damage (assuming that the passenger is well supported in the supine position), regardless of the peak deceleration. Since it is envisioned that manned capsules can be brought to Earth at velocities below 80 ft/sec, this prediction implies no necessity for shock mounting the man. One purpose of the impact test program reported herein was therefore to investigate the existence of this velocity asymptote.

Mice were selected because of their convenience as test subjects, compared with humans. Furthermore, estimates of their impact resistance showed behavior qualitatively similar to humans; so that the mouse impact program results could be expected to provide a valuable clue to the manner in which a human impact program should be conducted. The purpose of the program was therefore to determine whether an impact velocity asymptote exists for mice, to

establish the impact resistance curve for mice at velocities above the asymptotic value, and to reach some general conclusions as to human impact resistance.

THE IMPACT STRENGTH OF A STRUCTURE

A form of presentation of shock strength data, originally developed at the U.S. Naval Ordnance Laboratory[2] to describe the performance of inertia mechanisms, appears most suitable for portraying response to single acceleration-time pulses (single versus periodic forcing functions). The data is gathered by subjecting identical specimens to acceleration-time pulses of variable duration and amplitude. At each impact duration the minimum acceleration level which will cause damage is determined. Fig. 1 illustrates the manner in which duration affects the peak acceleration necessary to induce

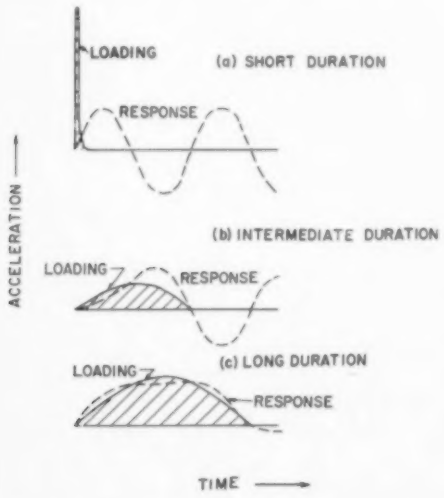


Fig. 1. Loading and response, half-sine pulses.

a given amplitude level of response. Note that the shorter the duration of loading, the greater the acceleration required for the same response.

Presentation of the data is accomplished by plotting the parameters ΔV and average acceleration of each pulse which just succeeded in damaging a specimen. The cross-hatched areas of Fig. 1 constitute the velocity change ΔV, while average acceleration is equal to velocity change divided by duration. Fig. 2 shows the "sensitivity curve" for a mass-spring system subjected to half-sine pulses. In application of the sensitivity curve, if the ΔV and average acceleration of an input pulse lies below or to the left of the curve the specimen experiences no damage, the damaging pulses falling above and to the right of the curve.

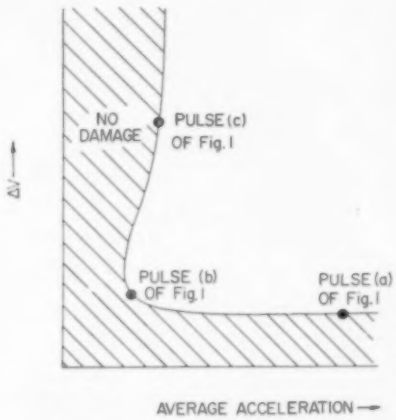


Fig. 2. Sensitivity curve for mass-spring system subjected to half-sine pulses.

Note that the sensitivity curve relates only to the parameters of the input function, no regard apparently being taken of response of the structure, coupling between mount and structure, etc. This is done purposely, so that the sensitivity curve may be used in any application where the input acceleration-time history is known. The effects of coupling and structural response are already reflected in the fact that damage was produced, and there is no practical profit in learning the details of the response of each portion of the mount–structure system. As long as one maintains the identical mounting system for identical specimens, the sensitivity curve should prove useful to predict damage or no-damage for any application which has known acceleration-time input characteristics.

If the specimens had undergone tests with triangular acceleration-time pulses, the results would be as shown in Fig. 3. Obviously the location of the vertical (or long-duration)

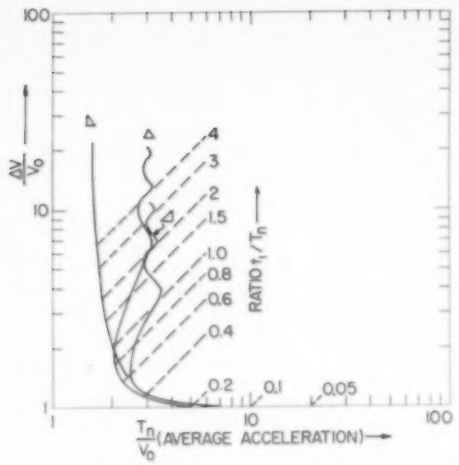


Fig. 3. Sensitivity curve for mass-spring system subjected to triangular pulses.

asymptote is a function of the shape of the input pulse. This implies the necessity of knowing more about the input pulse than ΔV and average acceleration. Note, however, that the horizontal (or short-duration) asymptote is independent of pulse shape, and may be specified by a unique value of velocity change. This constitutes a great simplification of some impact situations, the only specification being that the duration is short enough to be on the short-duration asymptote. Figs. 4–6 illustrate the physical situation for short-duration loading. These figures show that the low frequency specimen does not respond appreciably during the short time in which ΔV is accomplished, the peak response occurring long after the impact. The impact serves to set the mass of the specimen in motion (relative to the carriage), and the inertia of the moving mass produces stress and deformation. For this situation, the only significant impact parameter is ΔV.

Figures 2 and 3 apply to mass-spring systems with one degree of freedom. For more complex structures, with several modes of failure, the results might look like Fig. 7. In this case the cross-hatched area would represent the "safe" zone of operation.

Without entering into a detailed discussion of sensitivity curves,[2] for purposes of the present discussion it may be observed that the salient features of a sensitivity curve are the long-duration asymptote (for structures, this asymptote is determined by centrifuge test) and the short-duration asymptote. In particular, the existence of the short-duration asymptote is of utmost significance in impact considerations, since it implies that any single pulse with ΔV below the asymptote value will not be damaging, regardless of peak acceleration or duration.

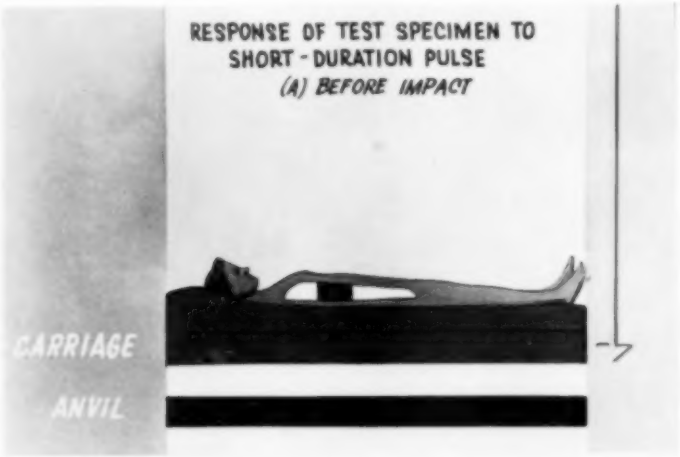


Fig. 4. Response of test specimen to short-duration pulse (A) before impact.

Fig. 5. Response of test specimen to short-duration pulse (B) immediately after impact.

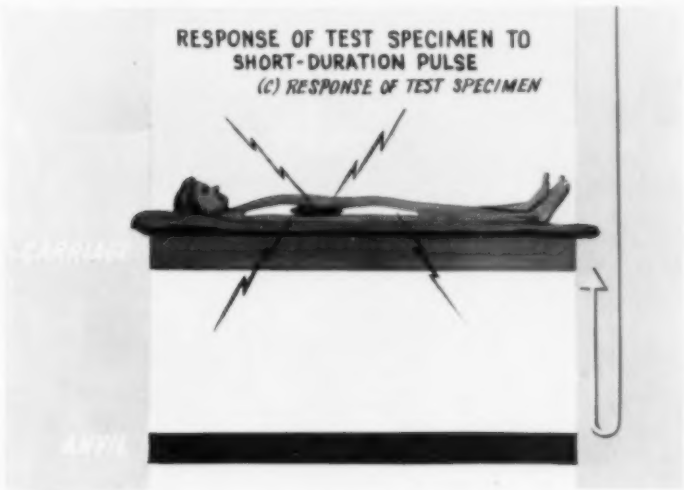


Fig. 6. Response of test specimen to short-duration pulse (C) response of test specimen.

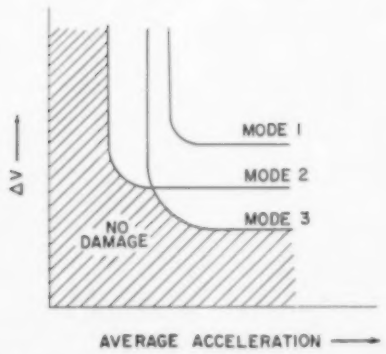


Fig. 7. Sensitivity curve for system with several modes of failure.

IMPACT STRENGTH OF THE HUMAN STRUCTURE

Figure 8 presents the experimental data on human impact strength in the form of a sensitivity curve.[1] The datum points indicated by square or triangular symbol[4] represent isolated accidents, while the circular symbol points[3] are only a few results of a planned experimental program. Unfortunately, however, correlation of the data as in Fig. 8 shows that the short-duration region (short-duration relative to human response to impact loading) is not touched by the sled tests[3] (circular symbols),

and estimates of the location of the short-duration asymptote must be based on the five square or triangular points shown. Conservatively drawing the short-duration asymptote near most of these latter points (since none of the points represent damage, the line could lie somewhat higher, possibly approaching the circular points: about 115 ft/sec) yields a lower limit of 80 ft/sec velocity change to effect damage to humans at any acceleration level. Stated simply, any impact at velocities lower than 80 ft/sec will not be harmful to humans.

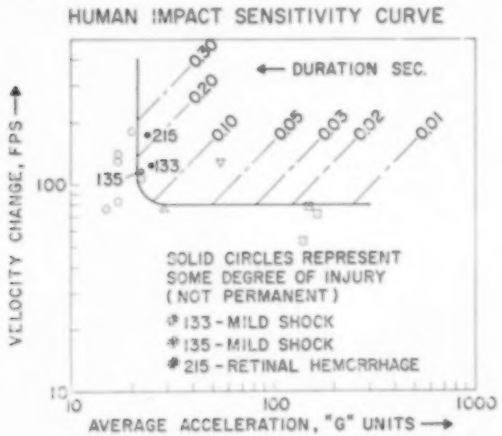


Fig. 8. Human impact sensitivity curve.

Note that the "long-duration" asymptote of Fig. 8 is considerably higher than the 10–15 G_{av} tolerance value usually quoted for centrifuge tests. This indicates several modes of failure (Fig. 7), the centrifuge mode of failure involving the shift of blood within the body over a long period of time. Also note that it was deemed premature to attempt to assess the effect of pulse shape on the long-duration asymptote, although Col. Stapp's conclusions regarding the importance of onset-of-acceleration are in agreement with the theoretical predictions of Fig. 3. That is, theoretically and experimentally, the acceleration pulse with a short rise time is more damaging than one with a longer rise time, other things being equal. For adequate design information in the long-duration region, it appears that more experimental data are needed using several pulse shapes. Pending the availability of this data, however, it is recommended that a theoretical approach[2] be employed for adjustment of pulse shape when necessary.

In general, the short duration asymptote is of greatest significance in impact situations. Unfortunately, however, the short duration asymptote of Fig. 8 is based on only five datum points. An experimental mouse impact program was therefore designed by General Electric's Missile and Space Vehicle Department specifically to produce statistical results in the short-duration region, in order to establish the existence of this asymptote on a firm basis.

DESIGN OF THE MOUSE IMPACT TEST

Figure 9 presents the test setup for impact testing of mice. At upper left is the drop-test carriage held up by an electromagnet, guided by a single cable which passes through the center

Fig. 9. Drop tester and instrumentation.

of the carriage. At lower left is the anvil which contains a strain-gage dynamometer for recording carriage deceleration, which is controlled by the size and shape of the stopping devices. In Fig. 9 the stopping devices are two lead cones,

Fig. 10. Test container and mice.

the deformation of which results in a gradual buildup of deceleration. Flat lead discs produce a somewhat rectangular acceleration-time pulse, while elastic devices such as springs produce half-sine acceleration-time pulses.

For each drop test, four mice were induced to crawl into clear plastic tubes (Fig. 10), the tubes loaded and secured in the carriage, and the carriage hoisted to the desired height. Impact velocity was controlled by the height of drop, with a maximum of about 47 ft/sec from the roof of the laboratory used for the tests reported herein. When the electromagnet was de-energized, the carriage fell under 1 g and, on impact, decelerated to zero velocity in a manner controlled by the stopping device. The deceleration-time pattern was recorded by the instrumentation shown at the right of Fig. 9. See Fig. 11 for typical records.

Besides controlling the impact parameters, it

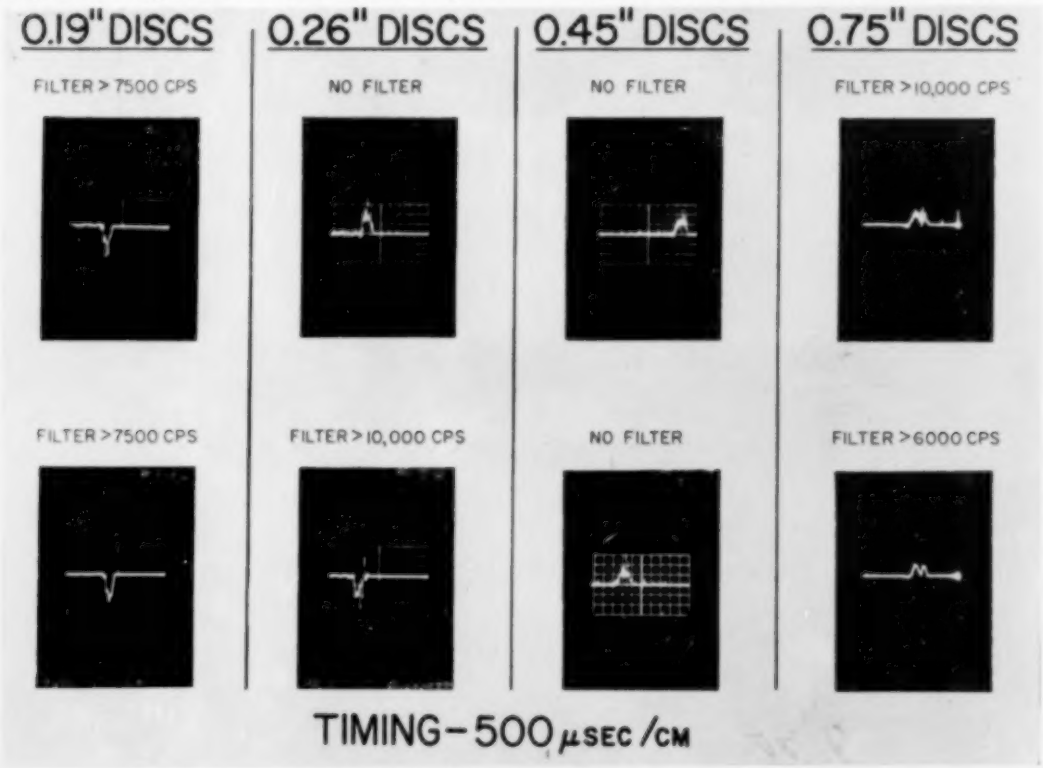


Fig. 11. Typical acceleration-time records.

was necessary to exercise careful control over the test specimens in order to avoid introduction of extraneous variables. Some complications were avoided by using one breed of mice (black mice, strain C-57, Jackson Memorial Laboratory, Bar Harbor, Maine), male, 4–6 weeks old. Initial orientation in the tube was belly down, recording being made of any cases of dead specimens being found in other orientations. The manner of supporting the mice could not be made uniform, due to weight and size variations in the mouse population. This variable was treated by using one tube size and determining statistically the significance of mouse weight.

Since it was not considered reasonable to find the threshold values of acceleration for inducing damage, which would require repeated impacts of each specimen, the tests were conducted on a go or no-go basis. As the g or ΔV level was increased the percentage of mice killed at each test condition increased, and the results were plotted as per cent killed versus ΔV. Generally, values of ΔV were increased in a stepwise manner using a particular stopping device. For most stopping devices as ΔV increased there was a concomitant increase in average acceleration, the technique not permitting variations in ΔV at constant average acceleration.

EXPERIMENTAL RESULTS

The initial data obtained in this program are displayed in Fig. 12, where per cent mortality is plotted against velocity change. For the different stopping devices, the curves show varying amounts of skewing, or departure from a symmetrical S-shape. An explanation for this is lacking at present, although multiple modes of failure may perhaps contribute. For example, Libber[5] has found several different thresholds of injury for rats subjected to accelerative forces of short duration.

The definition of damage has proved simpler than expected—the mice either died immediately after the test or appeared to recover completely. Thus the testing procedure has been on a go or no-go basis in practice. As a matter of fact, the mice which survived one test condition at a given mortality rate (50 per cent, for example) would show an even lower mortality rate (perhaps 20 per cent) on being re-tested at the same test condition. Thus the indications were that the first test served to weed out the specimens which were inherently less tolerant to impact. Whether the preconditioning introduced by the first test weakened or strengthened the survivors was not determined.

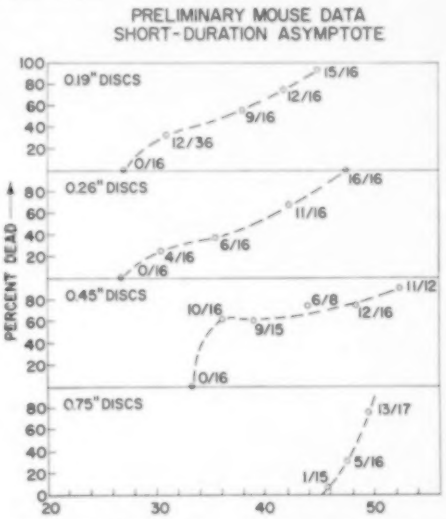


Fig. 12. Preliminary mouse data, short-duration asymptote.

Pathological examinations of selected animals have failed to show lesions which can be established as the cause of death. A number of animals have shown pulmonary hemorrhages (Fig. 13), in some cases sufficient to cause death, but this has not been a systematic finding. Upon removal from their tubes many animals showed convulsive behavior and slow deep respiration, but no central nervous system lesions have been identified on gross or microscopic examination. Recently electrocardiographic records (Fig. 14) have been obtained during free fall and during and after impact. No abnormalities or changes in heart rate have been observed during weightlessness. Bradycardia is a uniform finding immediately following impact.

No dependence of impact tolerance on weight has been found. This implies that the tube size is not critical and that the weight of the mouse is not, in itself, a significant parameter.

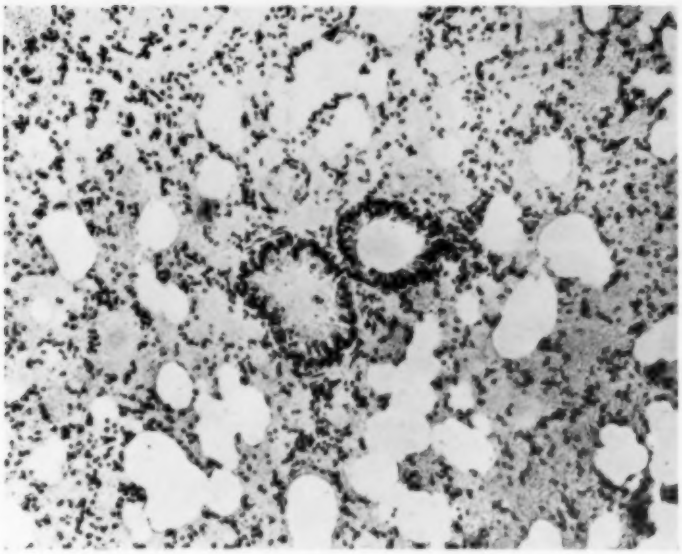


Fig. 13. Lung damage.

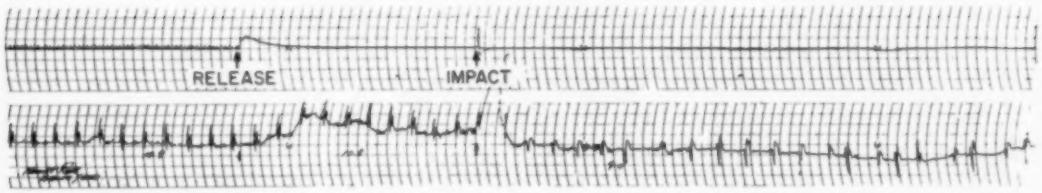


Fig. 14. Mouse electrocardiograph record.

The data of Fig. 12 are replotted in Fig. 15 to show the "mouse sensitivity curve". Mortalities of 1, 50 and 99 per cent are shown, together with an estimated long-duration asymptote.

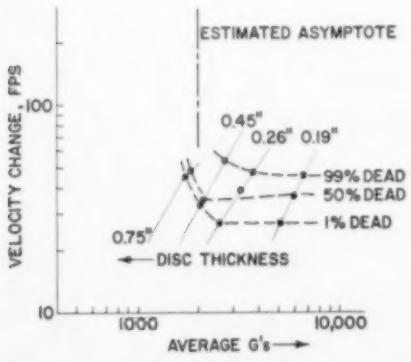


Fig. 15. Sensitivity curve for mice, preliminary data.

These data establish the existence of a short-duration velocity asymptote for mice, from the 2000 g long-duration asymptote up to about 10,000 g.

CONCLUSIONS

Although much work remains to be done in obtaining statistical data on the mouse impact sensitivity curve in the long-duration region where pulse shape assumes significance, the main purpose of the program has already been fulfilled. It is established by controlled experimentation that one mammal, the mouse, has an impact tolerance which can be described by the single parameter ΔV in the short-duration region. For the mouse, this may be put otherwise by stating that more than about 30 ft/sec velocity change is dangerous, regardless of peak acceleration or shape of pulse.

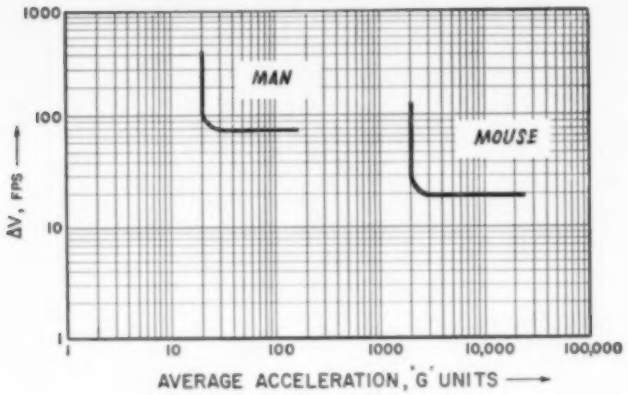


Fig. 16. Impact sensitivity curves.

Figure 16 shows the comparison between sensitivity curves for man and mouse. The difference between long-duration asymptotes conforms with established centrifuge data for two animals of different sizes. On the other hand, the fact that the velocity asymptote for mice is so much below that for man is somewhat unexpected, in view of the fact that similarly shaped inanimate structures of different sizes built of the same materials have the same short-duration asymptote values. The explanation perhaps lies in the difference in internal structures between man and mouse and in the different modes of failure. Data on an animal of a size between man and mouse would be of interest here.

Extension of this research to larger animals and finally to man should establish the human impact sensitivity curve firmly. It may then be possible to state that crashes are tolerable to well-supported humans at less than 100 ft/sec (or 68 m.p.h., or a free fall from 180 ft), regardless of peak deceleration. Information such as this will be of great significance in extra-terrestrial applications as well as such mundane situations as automobile and plane accidents.

Nomenclature

g	Acceleration of gravity, 32·2 ft/sec^2
G_{AV}	Time-average acceleration, g units
H	Height of drop of carriage under 1 g
t_1	Duration of impact
T_n	Natural period of vibration
V	Striking velocity
V_0	Velocity change required to produce a given amplitude of response
ΔV	Velocity change on impact

REFERENCES

1. M. KORNHAUSER, *Impact Protection for the Human Structure* in *Proc. Amer. Astronaut. Soc., W. Reg. Meet.*, August 1958, Palo Alto, Calif.
2. M. KORNHAUSER, *J. Appl. Mech.* **21**, — (1954).
3. J. P. STAPP, *Human Exposures to Linear Deceleration*, part 2 of *The Forward-Facing Position and the Development of a Crash Harness*, WADC AF Tech. Rep. 5915 (Dec. 1951).
4. H. DEHAVEN, *Mechanical Analysis of Survival in Falls from Heights of Fifty to One Hundred and Fifty Feet* in *War Medicine* **2**, 586 (1942).
5. L. M. LIBBER, *J. Avia. Med.* **28**, 166 (1957).

HUMAN ENGINEERING OR ENGINEERING OF THE HUMAN—WHICH?

ROBERT S. POGRUND
Space Technology Laboratories, Inc.

Abstract—The degree of compromise between the human engineer and the physiologist is a function of space flight duration. Short trips will permit the designing of the space capsule and its internal environment with a minimal degree of comfort to the human passenger. Medium length trips of interplanetary magnitude will require a duplicate of man's comfortable terrestrial environment, if he is to operate optimally as a scientific observer, capable of high performance levels. Extended interstellar space travel may not require the presence of incessant human intelligence, so that aids for ultimate survival are drawn from considerations of pharmacologic agents, hypothermia, adaptation and Yoga.

Our Space Program has evolved along a triphasic approach: (1) Technology within the physical sciences and rocket industry toward the design of space vehicular systems, including orbiting satellites and space probes, as well as the propulsive systems for placing them at their destinations; (2) application of the physical sciences toward evaluating and appraising the quality and quantity of phenomena existing in space, and these have included the investigations of radiation, atmospheric density, nature of the chemical atmosphere as space equivalence is approached, temperature changes, gravitational changes, meteor streams, placing telescopes and cameras in planetary orbit or landings in order to study the composition of other planetary atmospheres and geography, and so forth; (3) development of the Space Medicine Program in order to assess the above physical phenomena as to their posing obvious threats to the viability of man as the ultimate component of a space vehicular system.

This program has resulted in a vast interdisciplinary effort amongst engineers, physical scientists and medical or life scientists, and it is reasonably certain that while knowledge of the space environment is still incomplete, no important barriers stand in the way of unmanned flights. No such definite statement can be made where manned flight is concerned, partly because the requirements for human survival are much more severe than for instruments, and partly because instrument flights can be one way while manned flights must be round trips.

Space medical scientists have well defined the stresses expected to be imposed upon man in space, but opinion is somewhat divided as to the ability of man to withstand these stresses when defined as a function of operational time. The question arises as to whether or not the man should be completely safeguarded or are compromises permissible, and if so, how far can we go in these compromises, and what should be the degree and duration of these stresses to enable man to survive them?

The term "human engineering" has been coined as a discipline describing the efforts of engineers to protect man in space from the deleterious effects of these stresses, and these efforts have been geared toward the duration of the space trips. One of the main specific examples among the multitudinous tasks of the human engineer, has been the obvious recommendation for a hermetically sealed cabin with the creation of a completely artificial environment. For short trips, such as anticipated for Project Mercury, it has been considered feasible to carry along all the oxygen, chemically absorb all carbon dioxide by lithium hydroxide, and use evaporative cooling, carry all water and food, and store or discard waste material. A mission of medium duration of several weeks, carrying one or more passengers, may include regenerative absorbers for carbon dioxide controls, cryogenic oxygen

storage, vapor cycle cooling equipment, ashing or dehydrating waste and food, and regeneration of water from urine. An advanced type power supply will have to be included. A long mission of several months will necessitate the regeneration of oxygen from biologically active algae systems, which also absorb carbon dioxide as well as serve as food. Waste would be cycled back into the algae system to be used as a culture nutrient, while the only water will be regenerated from urine and perspiration.

Much work has gone into rendering these systems as practical as the theoretical considerations appear to promise, and progress is in evidence. Protection from acceleration of g forces, thermal stresses, noise stresses, radiation stresses, psychophysiological stresses, operational tasks, body cleanliness, exercise and movement, consideration of man's day–night cycles or diurnal physiological rhythms, the long-range effects of the subgravity state, possible disorientation, and many other considerations have been and are being investigated. However, much of this work is approached from investigating a particular stress as a single entity without due consideration to the interactions and synergism of combined effects. Simulation of these stresses at ground level is admittedly difficult, but, as long as algae systems are contemplated to meet man's metabolic requirements during space travel, then a space simulator of comparable weight and volume dimensions to that of a manned cabin in space flight, should be provided with such an algae system plus as many other above-mentioned parameters, in combination or singly, as possible. Human subjects will have to be exposed to such a space simulator over periods of time that are realistic for contemplated durations of interplanetary journeys. These are long term considerations, and undoubtedly, the space medical program will once more reveal the inevitable lag behind technological advances of rocketry and the development of more exotic means of propulsion, but the sooner such a program is initiated, the shorter their lag period, whereby man in space will be ready to assume the responsibility of accepting the advances in rocketry and utilize the physical capacities of the system as far as his physiological limitations will permit.

ENGINEERING OF THE HUMAN

The total human risk factor for manned space flight is difficult to delineate at the present time. As already indicated, the risk is in direct proportion to the duration of the space mission and the resultant complexity of the system engineered for man's protection and survival. Because of this complexity, consideration of the mathematical reliability of man's survival becomes significant. The first phase of space travel, that of exposing man to the space equivalent conditions of the aeropause, as planned for Project Mercury, is of imminent reality, since reliability of his return is excellent. On the other hand, the reliability of man's return from an interplanetary mission is low because of inadequate knowledge in the present state-of-the-art. Can the human engineer and the space medical scientist meet half way by engineering the human, and thereby, increasing the reliability of a successful interplanetary mission? Several approaches are possible, and these have been considered from studies of pharmacological aids, applications of hypothermia, adaptation techniques and Yoga.

I. Pharmacological Methods

A. *Radiation Protection**

The broad spectrum of compounds tested for protective activity ranges from simple gases and inorganic as well as organic compounds to complex biological materials, hormones and dietary factors, all exerting their protective effect through varying types of mechanisms. These materials cannot be injected after irradiation, but must be present in the body during the exposure to the radiation.

In general, one classification of compounds which exert their protective effect by lowering tissue oxygen tension is shown in Table 1.

* Excellent reviews of this topic may be found in the published volume: *Second United Nations International Conference on the Peaceful Uses of Atomic Energy*, Vol. 23, *Experience in Radiological Protection*, United Nations, Geneva (1958).

Table 1. Drugs for Increasing Tolerance to Radiation

Action by lowering tissue oxygen tension at subtoxic dosage levels	
Cyanide Malononitrile p-Aminopropriophenone Sodium nitrite	Pitressin Epinephrine Carbon monoxide
Sulfhydryl compounds	
$HS-CH_2-CH(NH_2-CO_2)(COOH) \rightarrow HS-CH_2-CH_2-NH_2$ Cysteine	Cysteamine
Glutathione	
$HOOC-CH(NH_2)-CH_2-CH_2-CO$ Glutamic acid	$-NH-C(H)(CH_2SH)-CO$ Cysteine $-NH-CH_2-COOH$ Glycine
Aminoalkylisothioureas and mercaptoalkylguanidines	
S, 2-Aminoethylisothiourea dihydrobromide (AET) S, 3-Aminopropylisothiourea (APT)	

Dowdy *et al.*[1] have previously shown the decrease in radiation mortality as an effect of severe hypoxia. He found 100 per cent survival at 30 days for rats irradiated in 5 per cent oxygen as compared with 100 per cent mortality for the controls and an $LD_{50}/30$ days of 1200–1400 r. The various agents, shown in Table 1, appear to exert their prophylactic activity against the damaging effects of ionizing radiation by way of a more indirect way of inducing a similar tissue hypoxia.[2–13]

The sulfhydryl compounds, including cysteine

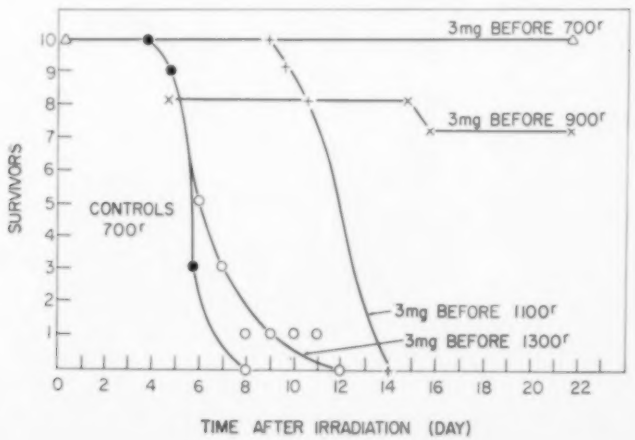

3mg BEFORE 700r
3mg BEFORE 900r
CONTROLS
700r
3mg BEFORE 1100r
3mg BEFORE 1300r
SURVIVORS
TIME AFTER IRRADIATION (DAY)
0 2 4 6 8 10 12 14 16 18 20 22
1 2 3 4 5 6 7 8 9 10

Fig. 1. Radiobiological protective action of cysteamine in mice. Intraperitoneal injection of 3 mg cysteamine immediately before irradiation with various doses of X-rays. [From Bacq and Alexander[15]]

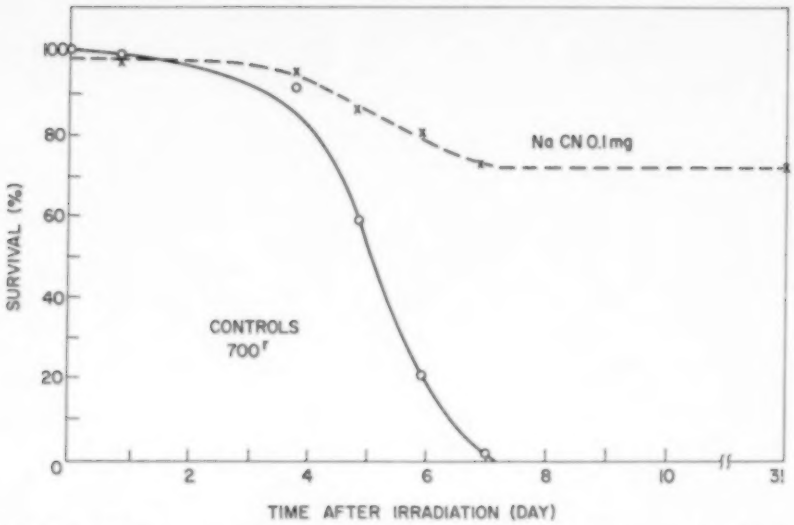


Fig. 2. Radiobiological protective action of cyanide in mice. Survival of control irradiated mice (700 r X-rays, 250 kV) and of cyanide-injected animals similarly irradiated. Intraperitoneal injection immediately before irradiation. [From Bacq and Alexander[15]]

and cysteamine, have been extensively studied for their protective action against radiation effects ever since the work of Patt,[7] who found an elevation of the $LD_{50}/30$ day radiation dosage of 765–1120 r at the maximally tolerated dose of 1200 mg/kg. Cysteamine has been found to be considerably more effective, producing a similar degree of protection, but at a much smaller dose of only 150 mg/kg.[8]

Glutathione[10–11] was found to raise the $LD_{50}/30$ days from 740–950 r after injecting 4 mg per mouse, while the protective effect was found mainly to reside in enhancing the recuperative power of the hemopoietic system.

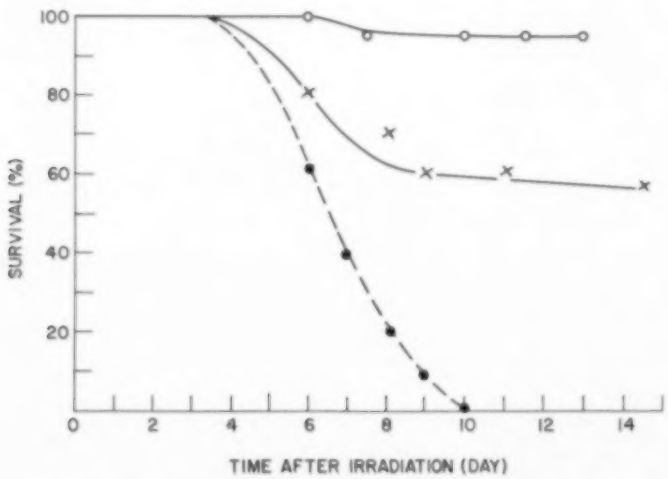


Fig. 3. Radiobiological protective action of glutathione and cyanide.
● Control mice (700 r)
× 2 mg glutathione injected
○ 2 mg glutathione + 0·1 mg NaCN injected. [From Bacq and Alexander (15)]

The aminoalkylisothioureas and mercaptoalkylguanidines[12–14] were demonstrated to be about 3 times more active than cysteamine, increasing the radiation $LD_{50}/30$ days from 625–1300 r.

Figures 1–3 show in graphical form the extent of increased $LD_{50}/30$ days radiation protection afforded by several of these compounds producing tissue hypoxia.[15]

Table 2 indicates a variety of miscellaneous compounds which have been reported to exert radiation protective actions in varying degrees of activity.[16–26]

Table 2. Drugs for Increasing Tolerance to Radiation

Miscellaneous compounds	
Cobalt	Somatotropic hormone
Magnesium	Glucose
Methyl linoleate	Colchicine
Chlorpromazine	Methimazole
Thiouracil	3-Aminotriazole
Cortisone with chlorpromazine	Bacterial endotoxins
Hexamethonium with cortisone	Homologous serum or plasma
Dibenamine	Infusion of living bone marrow suspension

described afford a protective action primarily through their effect on the hemopoietic system present within the bone marrow. The use of the term "protection" must not mislead one to believe that all three systems will be protected from the lethally damaging influence of ionizing radiation. From the material presented, a protective "factor of 2" may be maximally ascribed to some of the more efficient compounds, so that the tolerance to a given radiation dose may be increased from $LD_{50}/30$ days to say an $LD_{50}/60$ days. Or, the increased tolerance for a given dosage of 600 r, for example, producing

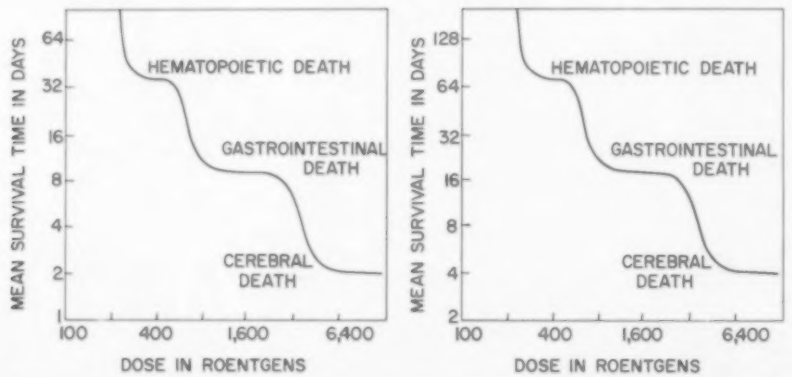


Fig. 4. (a) Acute effects of gamma radiation in primates. Dose–survival time relationship predicted for man. [From Gerstner[27]]. (b) Shift in curve to show possible extent of radiation protection by drugs. Dose–survival time relationship predicted for man under influence of drugs.

Exposure of the total body to penetrating radiation kills animals by at least one of three general mechanisms—injury to the bone marrow, to the gastro-intestinal tract or to the central nervous system,[27] as shown in Fig. 4. It is well to stress the point that the various compounds an $LD_{50}/30$ days, may be defined in terms of withstanding 1200 r for the same LD_{50}. The curve, shown in Fig. 4, may thus be shifted to include such protective action within the necessary boundaries of survival. Consequently, a radiation dosage that would normally be

lethal by its effect on the intestines, might now be tolerated for a longer time, under the influence of protective agents, while that same dose would still be lethal by its effect on the bone marrow. If, by definition, it is conceded that certain chemicals increase the life span of organisms exposed to an intolerable dose of ionizing radiation, then these chemical agents exert a protective action for that criterion. If a space astronaut should enter a radiation field where 400 r/hr are present, the injection of a chemical agent would be of little value if the astronaut were allowed the false security of dwelling for 2 hr instead of 1 hr, if his inevitable death from bone marrow damage occurred in 30 days. On the other hand, such degrees of chemical protection may be useful to the engineer faced with the problem of providing 6 in. thickness of lead as shielding for 99 per cent protection to the human, compared with 1 in. thickness of lead affording 90 per cent protection. The use of a chemical agent to provide the extra 9 per cent protection in order to eliminate the complex problem of an extra 5 in. thickness of lead, would be a clear example of perhaps a convenient compromise.

The matter of preventing this death by the injection of living bone marrow suspension after exposure to radiation, provides a different approach to protection. This method actually replaces a damaged bone marrow incapable of manufacturing platelets and erythrocytes, thus producing hemorrhage and anemia. The inected bone marrow grows in the irradiated marrow and soon forms new granulocytes.

At higher levels of radiation, where neither the previously described protective chemical agents nor the injected bone marrow are capable of permitting recuperation due to the severe intestinal damage, survival might still be prolonged by replacement of fluid and electrolytes.[28] No agent or procedure has yet been found to counteract the irreversible damage to the central nervous system resulting from exposure to very high dosage of radiation. Although hypothermia has been known to reduce the lethal effects of radiation in newborn rats and mice,[29–30] it is not certain that "deep-freeze" hypothermia can entirely prevent the tissue damage imposed by high radiation levels. Conclusive results on such an approach await a technique for long duration freezing and thawing followed by normal restoration of physiological function and prolonged survival.

B. *Drugs Affecting Metabolism*

Some of the drugs that might be of conceivable use to a space astronaut are listed in Table 3. These are compounds which exert their effect through the alteration of basic metabolic processes, and are of such general use in clinical practice that their properties can be found in any complete textbook on pharmacology.[31]

Table 3. Use of Drugs Affecting Metabolism

Chlorpromazine	Glucose
Thiouracil	Chlorpromazine and cortisone
Methimazole	Hexamethonium and cortisone
Glutathione	Dibenamine

Chlorpromazine[32] has been known to lower body temperature, and thereby decrease oxidative enzymatic metabolism. Such a pharmacologic tool would be useful if the oxygen supply within the sealed cabin has been diminished, for any reason, or if the temperature in the internal environment has been lowered to the point of uncomfortable constant shivering. The drug would then offer some degree of protection by decreasing the requirements of the body for oxygen through its body temperature lowering property. This drug apparently exerts a central action that abolishes the reflex impulses to the skeletal muscles in the same way as does any form of deep anaesthesia. This reflex blocking produces a lowered body temperature, since the skeletal muscles are a primary source of heat production, mediated by the hypothalmic centers. Thiouracil and Methimazole are compounds that are specific in lowering basal metabolic rates, should such needs arise as on extended space voyages where biological or chemical systems for generating oxygen might have become unreliable or malfunctioning. Glutathione is known to provide for greater resistance to

stress through enhancement of hydrocortisone activity on carbohydrate metabolism. The addition of cortisone to compounds, such as chlorpromazine or hexamethonium increases resistance to cold, as would the ingestion of glucose. Dibenamine is an adrenolytic which maintains a more homeostatic balance at the precapillary and capillary vascular level during exposure to stress. In general, compounds such as those mentioned would enhance the body's power of adjustment to a sudden or gradual change in environmental conditions.

C. *Drugs Affecting Motion Sickness*

The compounds listed in Table 4 are antihistaminics known to combat the effects of motion sickness. The need for such compounds might arise during the subgravity state and/or during exposure of the astronaut to low cycle vibrations or during spinning and tumbling, where disorientation, nausea, nystagamus or the oculogravid illusion may be evident as psychophysiological stresses.

Table 4. Use of Drugs for Motion Sickness

Antihistiminics
Diphenhydramine (Benadryl) Promethazine (Phenergan) Merezine Dramamine
Hyoscine

D. *Drugs Affecting the Excretory Systems*

Table 5 indicates the drugs that might be considered for use in the presence of digestive malfunction or of problems concerning micturition. Such usage might be indicated during exposure to the prolonged subgravity state, wherein muscle atony or even atrophy may arise as space voyages become extensive. The compounds listed primarily exert their effect on the smooth musculature, so that drugs, such as physostigmine, neostigmine and carbachol, may be used to stimulate contractions of the intestine, while also the last two enhance contractions of the bladder as well. Some space physiologists feel that the micturition reflex may be depressed under the influence of prolonged weightlessness, so that retention of urine could be a problem. Certain diuretic compounds may also be used to advantage, and these include caffeine, which decreases fluid reabsorption in the kidney tubules, theobromine, and acetozoleamide, which inhibits carbonic anhydrase and prevents reabsorption of the sodium ion. A situation may also arise where it might be necessary to decrease urine production, and concomitantly, the need for normal quantities of drinking water through fluid retention. Drugs, such as prednisone and certain amino acid mixtures, could meet this requirement.

Table 5. Use of Drugs for Excretory Problems

Stimulation of intestinal and/or urinary bladder contractions
Physostigmine (Eserine) Neostigmine Carbachol or carbaminoyl choline
Diuretic compounds
Caffeine Theobromine Acetozoleamide (Diamox)
Antidiuretic compounds
Prednisone Amino acids mixture

E. *Drugs Affecting the Cardiovascular System*

The dynamics of space flight necessarily will

affect the activity of the heart and peripheral blood circulation. Homeostatic balances of heart rate as influenced by blood pressure and states of vasomotion are based on the terrestrial level of 1 *g*. In the event that homeostatic balances are not restored in due time from the deleterious effects of acceleration, weightlessness or deceleration, then pharmacologic aids may again be indicated. Acceleration is expected to produce a high level of tachycardia in the space astronaut. As shown in Table 6, carbachol may be used to slow the heart in any state of prolonged paroxysmal tachycardia that might occur during both the exit and re-entry phase of space flight. Noradrenaline may be used where prolonged inactivity during the subgravity state results in smooth muscle atony within the walls of arterioles, venules and capillary networks, accompanied by a tendency toward low blood pressure. Also, during the subgravity state, vagal slowing of the heart rate could occur, so that the administration of adrenaline would increase tonicity while accelerating the heart rate and force of contraction. Adrenaline, by way of its property of decreasing capillary premeability, could also be used where circulatory stasis may result in the formation of petechial hemorrhages. Ephedrine has a similar action as adrenaline, but at a less intense level of response.

Table 6. Use of Drugs for Cardiovascular Problems

Cardiac decelerator
Carbachol or carbaminoyl choline
Pressor agents and cardiac accelerators
Noradrenaline
Adrenaline
Ephedrine

F. *Drugs Affecting Psychophysiological Control*[33]

Space flight of long duration will have psychological stresses imposed upon the human occupant as a result of isolation or possible other forms of sensory deprivations. Either tranquilizers or stimulants, depending on the circumstances, may be used to advantage to either minimize such psychological disturbances, or else render the astronaut to a higher peak of alertness or performance during an otherwise fatiguing phase of the mission. Table 7 lists the drugs that may be useful in these regards. The most important tranquilizers are reserpine and chlorpromazine.

Reserpine is considered to be one of the most potent of the Rauwolfia alkaloids, with respect to its sedative and antihypertensive effect accompanied by bradycardia. This effect is manifested through action on the vasomotor mechanism and resultant decreased peripheral vasoconstriction. In common with other drugs in this group, reserpine suppresses sympathetic activity at the hypothalamic level.

Table 7. Drugs for Psychophysiological Control

Tranquilizers
Reserpine (Serpasil)
Chlorpromazine (Thorazine)
Meprobamate (Miltown, Equanil)
Perphenazine (Trilafon)
Hydroxyzine (Atarax)
Ectylurea (Nostyn)
Stimulants
Amphetamine (Benzedrine)
Methylamphetamine
Caffeine
Metrazol

Chlorpromazine's unique pharmacologic action is through a preferential inhibition of those parts of the central nervous system which control vomiting, heat regulation, wakefulness, vasomotor tone, muscle tone and secretory activity of the anterior pituitary. This inhibition of subcortical centers, the reticular system, the thalamus and the hypothalamus of the brain, without significant action on the cortex, renders the drug especially useful in alleviation of anxiety, tension, agitation, vomiting from irradiation or motion sickness, and for the induction of hypothermia.

Examples of other tranquilizers include

meprobamate, perphenazine, hydroxzine and ectylurea. Table 7 also lists a few of the stimulants that may be used. Amphetamine and methylamphetamine and drugs that affect the central nervous system and remove fatigue, while producing a feeling of greater energy, alertness and wakefulness. Caffeine produces a similar response and also stimulates the respiratory center. Metrazol may have some use in improving the peripheral circulation through stimulation of the vasomotor centers of the brain.

The possible use of drugs as a protection from psychological stresses could become an important component for man's functional integrity during very extended space missions. The importance of the problem has been succinctly stated by Hauty,[34] wherein tests were conducted on subjects performing complex mental tasks for 30 consecutive hours. He observed, "following 12–15 hr of work and continuing up to the end of the 30 hr work period, all subjects, surprisingly enough, reported having experienced disturbances similar to those mentioned in another study. Their hallucinations and illusions ranged from simple and poorly defined phenomena such as, 'the instrument panel melting and dripping to the floor' to well-organized aberrations such as 'on several occasions the bank indicator showed a hippopotamus smiling at me'. Needless to say, these disturbances exerted quite an adverse effect upon proficiency. One subject, in fact, had to spend a good deal of time brushing away the little men that kept swinging on, and thereby obscuring, the air speed indicator." Hauty further commented on the 7-day simulated space flight study as follows: "It is significant that during the latter stages of the flight it was the subject's hostility which became a matter of increasing concern to the investigators. In fact, it reached the point of becoming the single conceivable reason for a premature termination of the flight."

II. Hypothermia

Hypothermia designates a state in which an organism is at a lower temperature than in the normothermic state. Homeotherms maintain normothermia through sensitive temperature regulation mechanisms present in the central and autonomic nervous systems, and this temperature is set at a relatively high level. A less apparent normothermia, in which the temperature is set and regulated at a low level, is the winter sleep of the hibernating animal, such as the bear. The extreme form of hypothermia is the deep frozen state. There can be no doubt about the interest and importance of the scientific aspects of this problem, but space does not allow a discussion of the various methods for inducing the hypothermic state at these various levels, nor can the degree of success be discussed. But suffice it to say that clinical medicine has now utilized the technique with remarkable success in cardiac and brain surgery, while, also, many clinical cases of accidentally induced deep hypothermia have been reported[35–36] with normal survival in some cases, and survival with amputation of limbs in other cases. The recent case of William Korpi,[37] aged 54, found unconscious in the frozen state of Drummond's Island in Lake Huron during sub-zero temperature, is one of the most unique cases, because an apparent high level of alcohol in his blood stream was the agent responsible for survival. Glycerol[38] has also been found to be protective during cooling of animals, while also Laborit's[39] mixture of atropine sulfate, phenobarbital, isocaine, phenergan and tubocurarine, is protective. By such use of chemical agents and the technique of extracorporeal circulation,[40] the reduction of body temperature to only 28°C, already reduces oxygen consumption to one-half of that occurring at 37°C. At 13°C, spontaneous cardiac arrest occurs, which can be counteracted by potassium citrate. By such methods, a surgical team at Duke University School of Medicine[41] was able to reduce the temperature of a surgical patient to 5°C, with a minimal oxygen consumption, followed by complete survival. Because of the complexity of the technique, research is now in progress, especially in England,[42–43] on deep freezing the whole intact organism, but with only moderate success, since survival is dependent upon duration of the induced

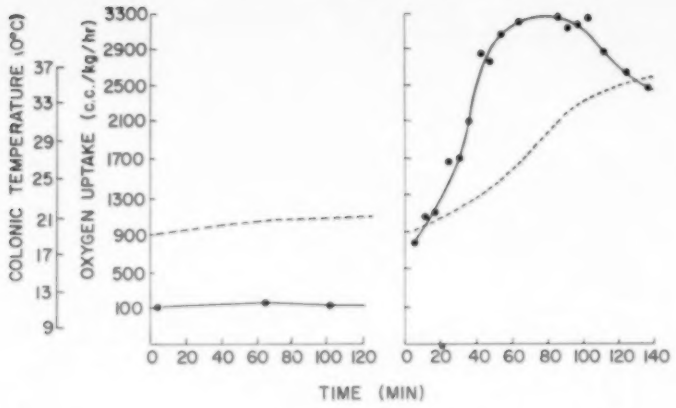


Fig. 5. The oxygen consumption and colonic temperature of two rats left at room temperature. (a) Previously cooled to + 2°C and rewarmed at 15°C. (b) Previously cooled to colonic temperature of 15°C.
● Oxygen consumption
- - - Colonic temperature

hypothermia. Fig. 5 graphically demonstrates the decreased oxygen consumption in a rat previously cooled to + 2°C in comparison to a rat cooled only to + 15°C. Fig. 6 shows the percentage survival of rats as a function of duration of freezing; thus indicating the degree of success with deep hypothermia.

It is obvious that considerations for applying the technique to man during space flight have no immediate bearing, since flights of short duration require man as a conscious, decision making and possibly navigating intelligent component of the man–machine system. Otherwise, the flight might, for all practical purposes, be considered as an unmanned vehicle. However, long range planning[44] should not exclude these possibilities for distant future manned space flights of interstellar system magnitudes, in which human intelligence might not be required while psychophysiological stress to the human,

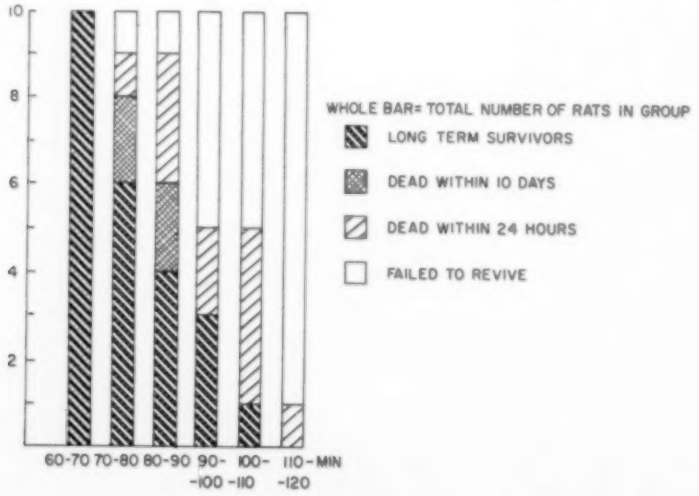


Fig. 6. Effect of the duration of suspended animation in rats maintained at 0 degrees of body temperature on the rate of recovery. [From Andjus and Smith[42]]

requirements for oxygen and biological damage from radiation[45] would be at very minimal levels during his function as merely a passenger. Such advanced thinking would possibly shorten the expected lag period between technological development of more sophisticated methods of propulsion and the ability of the human component to be contained as a passenger.

III. Adaptation to Stress

A common property of living cells is a responsiveness to the environment, and this interaction is an old but very important problem for biologists, who have considered this as one of the properties of living protoplasm. If an amoeba is placed in a dish of water, it can be observed to move toward certain chemicals and away from others. It can, therefore, react to the nature of its environment and even, perhaps, determine the concentration of substances around it, but to respond to the size of the dish seems a little too difficult for an amoeba. Bacteria are very good at adapting themselves to a new environment. They can grow in the presence of dyes which interfere with their ordinary mode of living, and during such processes of adaptation, new systems of enzymes are formed and new methods of synthesis are found to replace those which have been blocked by the dye. According to the metabolic theory, the cancer cell is an adaptation to new conditions, in which carcinogenic substances act by interfering with enzyme systems which the cell normally uses. These substances preferentially poison the cell which finds an alternate way of living in order to exist. Where colonies of bacteria are kept in an adverse environment, such as in the presence of an antibiotic, a change can occur, known as a mutation, which is transmitted to its progeny, thus forming a new resistant strain. Sometimes mutations are beneficial to the species, other times they are deleterious, as the possible mutations in man as a result of exposure to ionizing radiation.

Organisms respond to environmental change in different ways according to the time during which the environmental change persists and according to the magnitude of the stress. The adaptive variations of organisms may be genetically determined or they may be environmentally induced. Physiologists concerned with the physiology of man as an intact organism have paid most attention to adaptive variations induced by physical factors in the environment, such as temperature, radiation, oxygen tensions, carbon dioxide tensions, low atmospheric pressures, and so forth. Plant physiologists have been less concerned with environmentally induced variations in individual plants but more with analysis of the wealth of genetic variations as shown by strains adapted to a variety of environments. To use temperature as an example, a single enzymatic process that is particularly sensitive to temperature changes, may influence survival. If an essential metabolic compound that is involved in the process or reaction to temperature is known, then it should be possible to enhance survival by supplying it with the known metabolite, even if the organism is growing in a deleterious temperature. This concept has been worked out on temperature sensitive mutants of the red bread mold, *Neurospora crassa*, which grows quite well at temperatures of 35°–40°C (as shown in Fig. 7). A mutant was found that grows normally only up to 25°C, showing deterioration above that temperature. Apparently, the high temperature limits growth by controlling some essential reaction, corrected by the addition of B-vitamin riboflavin to the culture medium.[46]

It is most difficult to extrapolate such concepts to man's survival in adverse environments, such as might be encountered in space flights, wherein man might be adapted through training, to possess greater tolerances and thereby minimize his metabolic requirements. Possibilities reveal themselves only by what has already been observed in retrospect. The ability of man to acclimatize to low temperature, high temperature and low atmospheric pressure or hypoxia is well known, and such adaptations are usually accompanied by physiological and biochemical changes thoroughly discussed in the literature.[36, 47–58] These adaptations can be classified as those occurring in a matter of days, quite small by comparison with a life span.

The acclimatization to the high altitude of mountain climbing is one example. But, then this is to be compared with an adaptation in the classical evolutionary sense—a process involving large numbers of generations of a species and adaptability of man, which, given sufficient time, can increase his tolerance to specific stresses likely to be encountered during extended space flights. This aspect is presented only to show one avenue of approach toward compromising

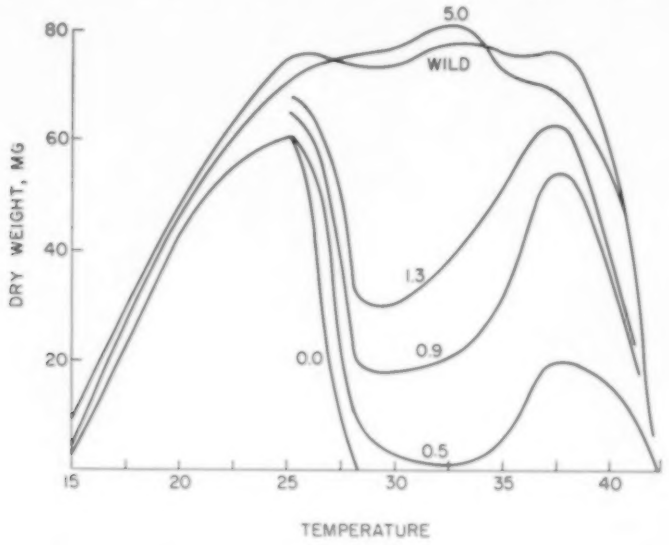


Fig. 7. Growth of wild type and temperature-sensitive riboflavin mutant of Neurospora at different temperatures and riboflavin concentrations. Riboflavin concentration in μg per 20 ml of culture. [From Mitchell and Houlahan[46]]

with a time course measured perhaps in millions of years. The very great differences in the time elements characteristic of each imply that there are actually very fundamental differences in the mechanisms which are operative in the different instances. An example of the long time element is the study carried out on the natives living in the Andean regions of Peru at elevations up to 16,500 ft. A most significant finding[59] is the close relationship between the level of altitude and degree of polycythemia (6,400,000/mm³), hemoglobin concentration (20 g/100 cm³ blood) and arterial oxygen saturation (80 per cent), as shown in Fig. 8. The cell volume, expressed in cm³/kg body weight, is about 60 per cent greater in the high altitude natives than in sea level residents, while the plasma volume is only 7 per cent lower.

It is not recommended that our astronaut be a Peruvian native because of his tolerance to hypoxia, but the examples are cited to show the with the engineer when considering the human factor as the limiting factor for manned space flight.

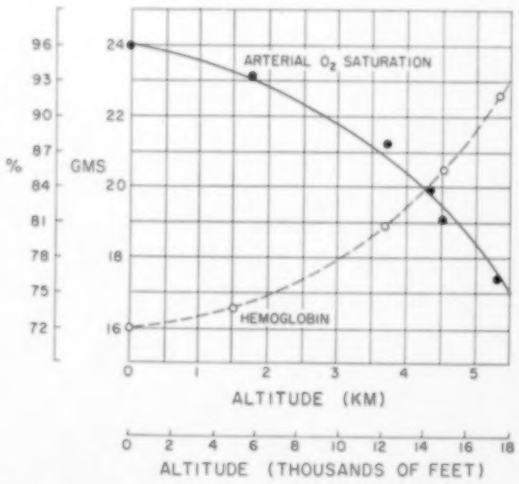


Fig. 8. Relationship between hemoglobin content, arterial oxygen saturation and altitude. [From Reynafarje[59]]

IV. The Application of Yoga to Space Medicine

This application is one of psychosomatic manifestation and may be considered still as part of an adaptation syndrome with specific physiological control through certain parts of the central and autonomic nervous system. The common denominator in these concepts is the effort toward homeostasis through the body's ability, to maintain itself as a "self centered unit" in the adjustment to stressful forces. The resultant "constancy" is under control or regulation by exceedingly complicated mechanisms via the sympathetic nervous system and/or the pituitary–adrenocortical axis and/or the hypothalamic – pituitary – adrenocortical axis. This control is purposive in that a most vital activity of mammalian bodies is the construction and maintenance of a protective shell—the internal environment, whose constancy of balance may be at a lower or higher maintenance level, depending upon the magnitude of the external and internal stimuli. All received stimuli are co-ordinated in a homeostatic center after the process of storage and selection, and emerge through effector organs, generally muscles and glands. These in turn act on the external world, and also react on the central nervous system through both effector and receptor organs of kinesthesia. The information received by kinesthetic organs is combined with his already accumulated store of stimuli to influence future action. These processes are projected on to certain portions of the brain as localized functions with respect to conscious and subconscious activities.

It is not the purpose of this account to discuss the various mechanisms or methodologies of Yogic practice, but the somato-psychic relationships resulting from such practice would bear considerations toward space flight for the human. It has already been reported that the Russians have borrowed six Indian Yogis to help train the candidate for space travel on breath and muscle control.[60] It is this author's opinion that a program of experimentation should be initiated in order to objectively evaluate this method as an aid in the solution of certain problems of human space flight. Such areas of uncertainty as the effects on the human of prolonged subgravity state lasting for days or months, potentially lends itself to resolution by the psychic control over somatic reactions. The concomitant varying degrees of sensory deprivation during the isolated and subgravity state could render a man inoperable in relatively short periods of time. The literature is reasonably filled with reports dealing with these effects of sensory deprivation in a man, ranging from obvious emotional instability, such as anger or depression, or near mental derangement.[34] Control of the environment by engineering is not the whole answer. The combination of the subgravity state and a lack of meaningful information input to the sensory receptor organs, may result in mental breakdown in most men in just a few hours or days.

It is doubtful if Western man can accomplish the degree of Yoga practice attained by the Indian Yogi trained through many centuries and generations, but moving in that direction could possibly enable the space astronaut to voluntarily minimize or at least decrease the effects of these impending stressful situations. The practice of Yoga can apparently control autonomic vegetative functions[61–63] to the extent of influencing oxidative metabolism, cardiac rhythm, breathing rate, the electroencephalograph, the electrocardiograph and the electromyograph of the practitioner. Such abilities to be turned on voluntarily would be obviously advantageous during very extended interplanetary and interstellar flights. The possibilities exist, even for Western man, as evidenced by the recent report[64] of a California man being able to induce progressive slowing of the pulse until there would be complete cessation of heart action, merely by sitting quietly and relaxing completely. During such time, his face became ashen-grey, typical of circulation failure. After a few seconds of this sensation, he could take a deep breath with the resumption of normal cardiac activity.

DISCUSSION AND CONCLUSIONS

Methods of adjusting man's homeostatic balances toward varying degrees of adverse

environmental stresses expected of space flight have been discussed from the standpoint of pharmacological means, hypothermia, adaptation and Yoga. Such considerations might lead to compromises beneficial to the human engineer engaged in the most important activity of creating the proper environment for the space astronaut. The engineer will be concerned with radiation, pressurization, ventilation, atmospheric composition, humidity, temperature, noise, vibration, illumination, volume of space cabin, information displays and many others, all concerned with maintaining the astronaut's efficiency as a thinking, decision-making human being.[65] The requirements are rigid, for the effort is essentially that of recreating his terrestrial home with all of its psychophysiological comforts. It would be premature to say at this time that these requirements can be fully met, since the only true and best test is the empirical one of sending a man into orbital flight for the arbitrarily designated duration and observe survival or decrements in performance. Such experimental work on human beings would not be tolerated in this country, so that animals are used first while humans are used for simulated tests on ground level. Obvious limitations are simulating the subgravity state either singly or in combination with other expected stress forces, as well as the continuous operation of a closed or partially closed cycle ecological system over realistic periods of time. It becomes clear that the feasibility of manned space travel is a function of the duration of the mission, and that human tolerances to various environmental stresses, singly or in combination, will be decremental as the mission becomes longer. Compromises with those individuals, concerned with engineering the environment within the space vehicle, can be made only for short periods of space flight, wherein the astronaut may be allowed to be harnessed in a pressure suit, while locomoting in a very narrow confine of a very limited volume contained in a space capsule of low ambient pressure. When this mission is prolonged to one of months or many years duration, the terrestrial environment, with the very narrow ranges of limitation,[65] must be duplicated, and this includes the provision for a team of astronauts able to perform and locomote within generous space volumes of the space vehicle, and able to wear comfortable clothing, all analagous to provisions made for occupants of a prolonged submerged submarine. Unless engineering technology develops the technique of erecting space stations at conveniently located orbits whereby the astronaut can navigate through interplanetary space by "space station island hopping" or by "lunar base hopping", then it would appear that the only compromises possible are those rendered by pharmacological, biophysical, adaptative and psychological aids. True, these methods may mean that we no longer have "man", as a thinking rational human being,[66] capable of making "precise scientific observations and constant monitoring of navigation and flight controls",[67] but the need for these capabilities may not always be present during an extended space flight. It is conceivable that manned space stations,[68] used as outposts for scientific laboratories as well as a conditioning area for adapting the space crew to an alien environment, in combination with the above aids for extended space flight might further ensure the success of an extended space flight mission.

The essence of the matter is to be futuristic as well as practical. True, for the initial attempts at man's orbital flight, he will have to possess his full mental faculties for various performance tasks over relatively short duration flights. But will this need be evident over the many months or years of interplanetary and interstellar flight of the future, wherein man, as we know him, could be either a subman or superman,[69] as determined by the need, for the sake of survival? Engineering of the human to meet the human engineer half-way should be considered now for the unpredictables of the future.

SUMMARY

(1) The complexity of engineering for the comfort of the human as an astronaut in a space vehicle has been emphasized. This complexity has been described as a function of mission duration, as well as a function of operational

requirements and performance capabilities expected of the human occupant.

(2) Where high levels of performance are not required, as during very extended space missions, aids toward "engineering" the human have been suggested with the aim toward compromising the rigid requirements assigned to the human engineer for space capsule design of internal environment.

(3) These aids have been drawn from (i) pharmacological concepts, wherein chemical agents are suggested for radiation protection, metabolic alterations, motion sickness, excretory problems, cardiovascular dynamics changes and enhanced resistance to psychophysiological stress; (ii) the application of hypothermia for increased tolerance to radiation, lowering metabolic oxidative requirements and also for psychophysiological problems; (iii) the provision for long term pre-training to allow for adaptation to specific or combined psychological and physiological stresses; and (iv) the use of Yogic practice for voluntary control of certain psychophysiological functions ordinarily considered to be under involuntary or autonomic control for Western man.

(4) These matters are discussed on the basis of future considerations for very extended manned space flight, to be balanced by present considerations of methods of approach and information retrieval from the academic literature. The ultimate goal is to minimize the lag existing for manned flight behind the progress of space technology.

REFERENCES

1. A. H. Dowdy, L. R. Bennett and S. M. Chastain, *Radiology* **55**, 879 (1950).
2. Z. M. Bacq and A. Herve, *Brit. J. Radiol.* **24**, 617 (1951).
3. J. B. Storer and J. M. Coon, *Proc. Soc. Experim. Biol. Med.* **74**, 202 (1950).
4. L. J. Cole, V. P. Bond and M. C. Fishler, *Science* **115**, 644 (1952).
5. J. L. Gray, E. J. Moulden, J. T. Tew and H. Jensen, *Proc. Soc. Exp. Biol., N.Y.* **79**, 384 (1953).
6. E. B. Konecci, W. F. Taylor and S. S. Wilks, *Radiation Res.* **3**, 157 (1955).
7. H. M. Patt, E. B. Tyree, R. L. Straube and D. E. Smith, *Science* **110**, 213 (1949).
8. R. Rugh, *Amer. J. Physiol.* **189**, 31 (1957).
9. D. G. Doherty, W. T. Burnett, Jr. and R. Shapira, *Radiation Res.* **7**, 13 (1957).
10. W. H. Chapman and E. P. Cronkite, *Proc. Soc. Exp. Biol., N.Y.* **75**, 318 (1950).
11. E. P. Cronkite, G. Brecher and W. H. Chapman, *Proc. Soc. Exp. Biol., N.Y.* **76**, 396 (1951).
12. R. Shapira, D. G. Doherty and W. T. Burnett, Jr., *Radiation Res.* **7**, 22 (1957).
13. D. G. Doherty and W. T. Burnett, Jr., *Proc. Soc. Exp. Biol., N.Y.* **89**, 312 (1955).
14. D. G. Doherty, R. Shapira and W. T. Burnett, Jr., *J. Amer. Chem. Soc.* **79**, 5667 (1957).
15. Z. M. Bacq and P. Alexander, *Fundamentals of Radiobiology*, Academic Press, New York (1956).
16. W. Parr, T. O'Neill and A. Krebs, *Science* **117**, 155 (1953).
17. H. C. Blount, *Radiobiology* **65**, 250 (1955).
18. H. J. Deuel, Jr., A. L. S. Cheng, G. D. Kryder and M. E. Bingemann, *Science* **117**, 254 (1953).
19. T. J. Haley, W. G. McCormick and E. F. McCulloch, *Proc. Soc. Exp. Biol., N.Y.* **88**, 475 (1955).
20. M. W. Spellman, F. E. Roth, L. Blank and C. W. Lillehei, *Cancer* **8**, 172 (1955).
21. W. W. Smith, I. M. Alderman and R. E. Gillespie, *Amer. J. Physiol.* **191**, 124 (1957).
22. W. W. Smith, *Science* **127**, 340 (1958).
23. A. N. Stroud, A. M. Brues and M. M. Summers, *J. Nat. Cancer Inst.* **15**, 1109 (1955).
24. R. N. Feinstein and S. Berliner, *Science* **125**, 936 (1957).
25. K. A. Porter, *Brit. J. Exp. Path.* **38**, 401 (1957).
26. P. Urso and C. C. Congdon, *Blood* **12**, 251 (1957).
27. H. B. Gerstner, *U.S. Forces Med. J.* **9**, 313 (1948).
28. R. A. Conard, E. P. Cronkite, G. Brecher and C. P. A. Strome, *J. Appl. Physiol.* **9**, 227 (1956).
29. S. Hornsey, *Nature, Lond.* **178**, 87 (1956).
30. J. B. Storer and L. H. Hempelmann, *Amer. J. Physiol.* **171**, 341 (1952).
31. J. H. Burn, *Lecture Notes on Pharmacology*, Blackwell Scientific Publications, Oxford (1958).
32. L. Chevillard, H. Giono and M. C. Laury, *C. R. Soc. Biol., Paris*, **152**, 1074 (1958).
33. A. L. Welsh, *Psychotherapeutic Drugs*, Thomas, Springfield (1958).
34. G. T. Hauty, *Air Univ. Quart. Rev.* **10**, 89 (1958).
35. R. J. Hock and B. G. Covino, *Sci. Amer.* **198**, 104 (1958).
36. H. Strughold, *Hypoxia, Hypoxidosis, Hypoxidation, Hibernation, Apparent Death and Suspended Animation*, Air Univ., USAF School of Av. Med., No. 56-14 (1955).
37. *Scope Weekly*, Vol. 4 (11 Feb. 1959).
38. A. U. Smith, *Proc. Roy. Soc. Med.* **47**, 57 (1954).
39. H. Laborit, *Réaction organique à l'Agression et Choc*, Masson, Paris (1952).
40. F. Gollan, *The Physiology of Cardiac Surgery* in *Hypothermia, Extracorporeal Circulation and Extracorporeal Cooling*, Thomas, Springfield (1958).
41. *Scope Weekly*, Vol. 4 (22 April 1959).

42. R. K. Andjus and A. U. Smith, *J. Physiol.* **128**, 446 (1955).
43. A. U. Smith, *Resuscitation of Frozen Mammals*, in *New Sci.* (30 Oct. 1958).
44. J. Lyman, *Some Aspects of Research Planning for the Psychophysiological Environment in Manned Space Travel*, presented at Symp. A.A.A.S. Meeting, Indianapolis (27 Dec. 1957.)
45. E. E. Godfroi in *Proc. 2nd Int. Conf. Peaceful Uses of Atomic Energy, Geneva*, Vol. 23, p. 76, United Nations, New York (1958).
46. H. K. Mitchell and M. B. Houlahan, *Amer. J. Bot.* **33**, 31 (1946).
47. C. H. Wyndham and J. F. Morrison, *J. Appl. Physiol.* **13**, 219 (1958).
48. R. B. Mefferd, Jr., H. B. Hale and H. H. Martens, *Nitrogen and Electrolyte Excretion of Rats Chronically Exposed to Adverse Environments*, Air Univ., USAF School Av. Med., No. 57-66 (1957).
49. H. B. Hale and R. B. Mefferd, Jr., *Amer. J. Physiol.* **195**, 739 (1958).
50. G. B. Sutherland, I. L. Trapani and D. H. Campbell, *J. Appl. Physiol.* **12**, 367 (1958).
51. I. H. Kaiser, J. N. Cummings, S. R. M. Reynolds and J. P. Marbarger, *J. Appl. Physiol.* **13**, 171 (1958).
52. R. B. Mefferd, Jr. and H. B. Hale, *Amer. J. Physiol.* **195**, 726 (1958).
53. E. Page, *Rev. Canad. Biol.* **16**, 269 (1957).
54. A. B. Carran, *Science* **129**, 784 (1959).
55. R. B. Mefferd, Jr. and H. B. Hale, *Amer. J. Physiol.* **195**, 735 (1958).
56. D. Criscuolo, H. B. Hale and R. B. Mefferd, Jr., *J. Appl. Physiol.* **13**, 353 (1958).
57. H. B. Hale and R. B. Mefferd, Jr., *Amer. J. Physiol.* **194**, 469 (1958).
58. P. Siekevitz, *The Man of the Future* in *Nation* (13 Sept. 1958.)
59. C. Reynafarje, *The Influence of High Altitude on Erythropoietic Activity* in *Homeostatic Mechanisms*, Brookhaven Symposia in Biol. No. 10 (1958).
60. *Periscoping the World* in *Newsweek* (18 May 1959).
61. K. T. Behanan, *Yoga, A Scientific Evaluation*, Dover Publications, New York (1937).
62. S. L. Vinekar, *Asanas in Every Day Life* in *Yoga-Mimamsa*, **7**, 33 (1957).
63. N. N. Das and H. Gastaut, *Electroencephal. Clin. Neurophysiol.*, Suppl. No. 6, p. 211 (1957).
64. *Scope Weekly*, Vol. 4, p. 3 (22 July 1959).
65. T. C. Helvey, *Missiles and Rockets* **5**, 68 (1959).
66. G. W. Hoover, *The Man Machine System in Space Vehicles*, presented at Amer. Physiol. Soc., Wash., D.C. (30 Dec. 1958).
67. H. A. Imus, *Human Factors in Probing Outer Space*, Preprint No. 59-1, Amer. Astro. Soc., Western Nat. Meeting (4 Aug. 1959).
68. K. Ehricke, *Astronautics*, **4**, 20 (1959).
69. J. Rostand, *Can Man Be Modified*, Basic Books, New York (1959).

COMPOSITION OF RADIATION TRAPPED IN THE GEOMAGNETIC FIELD AT ALTITUDES UP TO 1000 KILOMETERS

FRANCIS E. HOLLY
Air Force Special Weapons Center

and

RICHARD G. JOHNSON
Lockheed Missiles and Space Research Laboratory

Abstract—Several methods are available for making measurements above the Earth's atmosphere. One of these is a "parasite" package which rides "piggy-back" on ICBM developmental flights. This paper discusses one such system in use by the Air Force Special Weapons Center for making measurements in the region of the lower Van Allen radiation belt. One such experiment and results presently available are discussed. Data from one phase indicate that the radiation penetrating 30 mg/cm² is predominantly electrons.

I. INTRODUCTION

The Air Force Special Weapons Center has, for some time, been interested in measurements about the Earth's atmosphere. It is the purpose of this treatise to discuss one vehicle used in such measurements and one experiment, of several, made to determine the particulate composition of the lower Van Allen radiation belt.

II. VEHICLE

There are several types of systems which are suitable as instrument carriers in such a program, among these are: satellites, amphibiotic sounding rockets and the parasitic instrument package which uses some other rocket or missile as a carrier. The latter has been used successfully in several instances,[1–3] using ICBM developmental flights as carriers. The disadvantages of such a system are obvious, one cannot select the launch time or the trajectory and the failure probability is greater than with a developed sounding rocket system. Among the advantages is the "free ride" to altitude. Within limits all necessary weight is available; however, with the development of the more exotic sounding systems such as Javelin/Journeyman and Scout this is not an advantage singularly enjoyed by ICBM "parasite" vehicles.

The AFSWC pod, a self-contained system, was designed for use on the Thor-Able system but was actually used exclusively on Atlas systems, weighs approximately 50 lb and is an approximately rectangular shape (Fig. 1) with outside dimensions of 6 × 12 × 24 in. Mounting is accomplished with a four point suspension to brackets already provided on the missile. Two points are used for attachment, and two points serve as supports against longitudinal forces and contain ejection springs, each of which are compressed to approximately 50 lb force at the time of installation.

A one point attachment is desirable; however, in order to leave the central portion free from obstructions (as desired for placement of experiments) a two-point attachment must be utilized. Attachment is accomplished by means of two release bolts (Fig. 2). These bolts are presently piston-actuated, each piston being driven by a pair of squibs (one is sufficient) operating from a pre-set timer. The timers are started at lift-off by a pull-out lanyard system. The timers which are at the rear of the pod are connected in parallel. The timers and lanyard system are

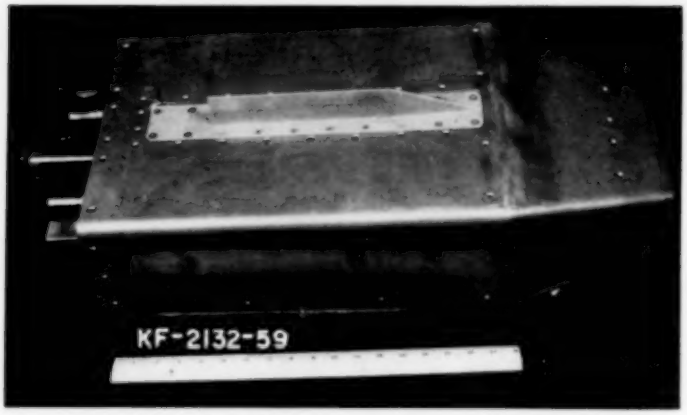


Fig. 1. AFSWC instrumentation package.

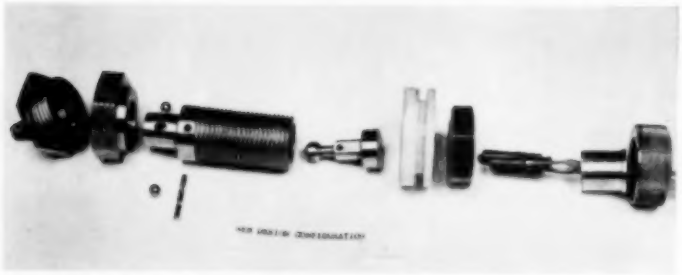


Fig. 2. Ejection and release bolt mechanism.

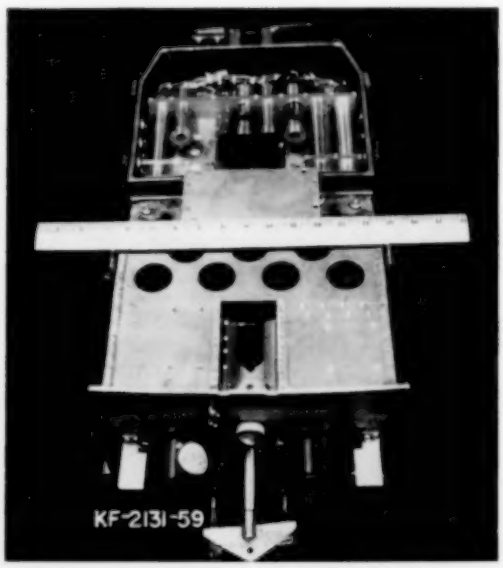


Fig. 3. AFSWC instrumentation package, bottom view.

shown in Figs. 1 and 3. Explosive bolts manufactured by Hercules Powder Company are under consideration for future programs.

Telemetry for the present system consists of a time multiplexed eight channel capability on a single FM carrier. The power output of the power amplifier is 10 W which is fed to a quarter wave antenna attached to the top of the package.

III. HISTORICAL BACKGROUND

Several measurements have been made in the region of the radiation belts first discovered by Van Allen which support the hypothesis by Stormer and others that charged particles may be trapped in the geomagnetic field. Several theories have been advanced about the origin of these belts but, as yet, relatively little is known about the charge, mass and energy spectrum of the trapped particles. The measurements

described herein and elsewhere[1–4] seem to support the theory that the inner zone consists dominantly of electrons from beta-decay of "splash-albedo" fast and slow neutrons (end-point energy 782 keV) and protons with energies up to several hundred MeV. Our measurements and those of Van Allen[5] indicate that the electron flux is at least three orders of magnitude greater than the proton flux.

IV. EXPERIMENTAL TECHNIQUE

Three instrumented packages (Figs. 1 and 3) were prepared each containing eight Geiger counters shielded and collimated in various ways (Fig. 4). Both side and end window counters were used, the end window variety being collimated to half angles of nearly 6° by cylindrical entrance apertures. Table 1 shows the various amounts of shielding which were used and the associated effective electron and proton energy thresholds. Numbers in parentheses indicate which counters were contained in each of the three packages.

In each of the packages at least one pair of counters were included which were identical in every respect (including orientation) except the entrance aperture of one was in a transverse magnetic field. These were counters *A* (2,3), *B* (1), *C* (3) and *E* (2). This magnetic field turns the effective collimator axis for electrons in the energy region of interest by more than 20°; the effective collimator axis for protons of energy great enough to penetrate the absorber is turned a negligible amount. In addition the overall electron sensitivity of these counters in an isotropic flux is lower because of reduced transmission through the collimator, while the proton sensitivity remains the same.

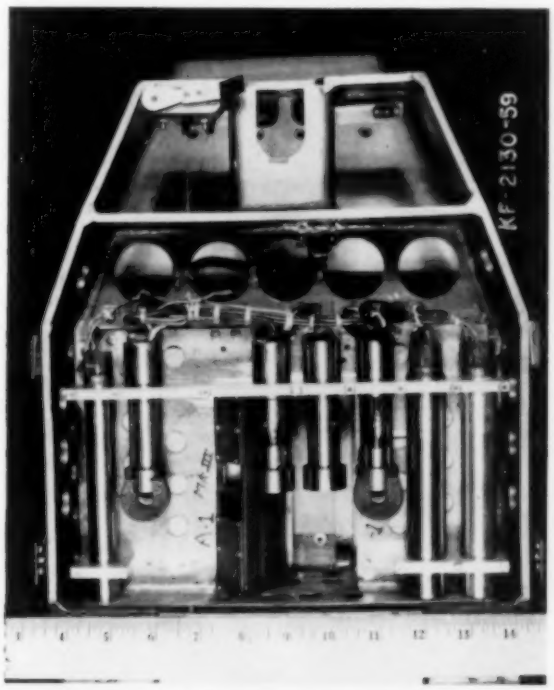


Fig. 4. AFSWC instrumentation package, close-up of Geiger.

The velocity vectors of trapped radiation at low altitudes have been shown to be confined predominantly to a plane.[6] This, plus the fact that the instrument package rotates or tumbles in flight permits the use of data from such identical pairs of counters to distinguish between electron and protons (or other heavy particles) as follows: If the z-axis is taken along

the transverse magnetic field, then the xy-plane contains the collimator axes, separated by more than 20° for electrons but coincident for protons. Two simple cases may be considered. First, if the axis of rotation lies in the plane of the flux and the z-axis is parallel to the axis of rotation, each counter will have a peak counting rate twice each revolution as it looks into the plane of the flux. The electron peak in the counter containing the magnet will occur approximately 20° out of phase with its twin counter but the proton peak occurs at the same time in both counters for electrons the peak rate in the counter containing the magnetic field will be smaller than in its twin because of reduced transmissions, while for protons the peak rates will be equal. Second, if the axis of rotation lies in the plane of the flux and the z-axis is perpendicular to the axis of rotation (or a line parallel to it), each counter will again have a peak rate twice each revolution (when xy-plane coincides with the plane of the flux), but the peak rates will occur at the same time in both counters for both electrons and protons; however, the ratio of peak rates in the two counters will still depend on whether the radiation is electrons or protons.

Utilizing the data from these pairs of counters as well as the range spectra provided by the other counters, one can determine a fairly good energy spectrum for trapped electrons. Assuming that the remaining particles are protons, a crude proton spectrum (for lower energies) may be determined.

V. RESULTS TO DATE

On February 7, 1959 and July 21, 1959, packages 1 and 2 were carried aloft and released.

Table 1. Counter Data

Counter	Shielding (mg/cmt²)	Electron Threshold (keV)	Proton Threshold (MeV)
A (1,2,3)	1·7 Mica	30	0·5
B (1)	30 Aluminum	160	3·8
C (3)	10 Aluminum	120	2·8
D (1,2,3)	30 Steel	190	4·6
E (2)	36 Aluminum	200	4·8
F (1,2,3)	150 Aluminum	460	10·0
G (1)	400 Aluminum	1000	16·5
H (2,3)	600 Aluminum	1500	23
I (1)	2000 Brass	4000	41
J (2,3)	3000 Brass	6000	52

At this writing data from only the first have been analysed. Fig. 5 shows the trajectory of the instrument package after release as well as the altitude variation of the peak counting rate in several of the counters, the curves being normalized to unity. Because of the altitude variation of the iso-count contours at several latitudes the trajectory achieved maximum penetration into the radiation after reaching apogee.

As shown in Table 1, this package contained counters of types *A*, *B*, *D*, *F*, *G* and *I*. Two counters of type *D* had a relatively large active area and thus were saturated at the points of interest; however, they were helpful in determining package orientation. The remaining six counters provided the data to be discussed. Fluxes as high as $2{\cdot}2 \times 10^5$ particles/cm²/sec $\pm$ 10 per cent and as low as 75 ± 6 particles/cm²/sec were encountered in counters *A* and *I* respectively.

Type *B* counters were used in this package as the matched pair, and during maximum penetration of the radiation data were obtained, at various times, which closely approximated the

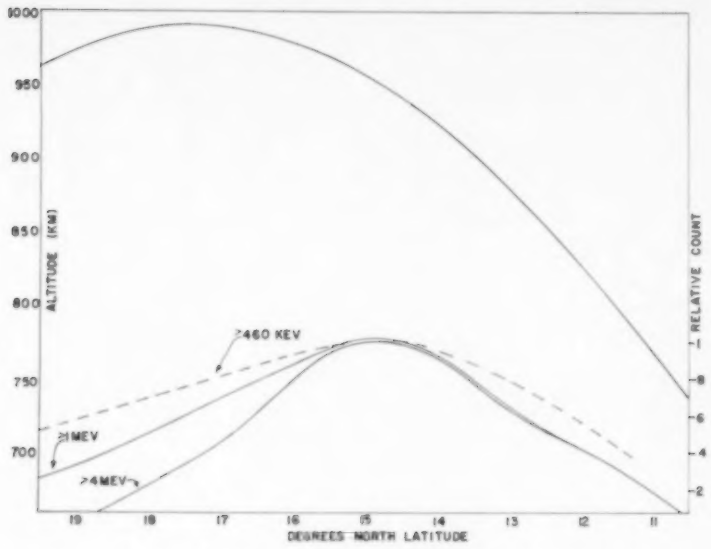


Fig. 5. Package trajectory and radiation profile.

idealized conditions described in the previous section. A quantitative analysis yielded an upper limit of 3 per cent (+ 5/− 3 per cent) for the proton contribution to the radiation penetrating the 30 mg/cm² absorber.

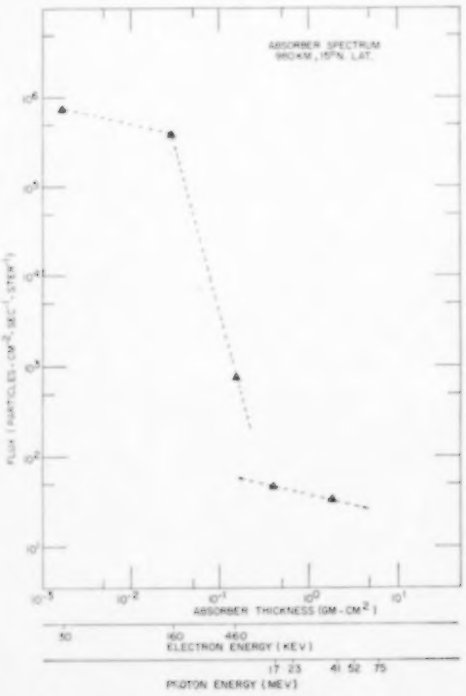


Fig. 6. Measured absorber spectrum (integral).

Calibrations were made with Tl^{204} and W^{185} sources, the combined spectra of which closely approximated the observed radiation. Fig. 6 shows the integral absorber spectrum of the observed radiation. *Less than 1 per cent of the radiation penetrated 150 mg/cm² of aluminum*; therefore, no estimate of proton contribution to the radiation can be made from this data, except in the 30–150 mg/cm² interval. However, a qualitative look at relative counting rates provided by our data and that of White[2] and Rosen[3] leads one to suspect that the radiation which penetrates 400 mg/cm² of aluminum is predominantly protons. This is not in conflict with the Earth neutron albedo concept.

Acknowledgments—The authors wish to thank L. Allen and J. A. Welch, Jr., of the Air Force Special Weapons Center, and R. D. Moffat, M. Walt and A. Meyerott of Lockheed Missiles and Space Research Laboratory for their assistance in instrument design and data interpretation. The authors also wish to express their appreciation to the Mathematics Branch of the Physics Division at Air Force Special Weapons Center for help in data reduction and machine calculations.

REFERENCES

1. F. E. Holly and R. G. Johnson, *Composition of Radiation Trapped the Geomagnetic Field at Altitudes up to 1000 Kilometers*, USAF-AFSWC-TN 59-15.
2. S. R. White and S. C. Freden, *Protons in the Earth's Magnetic Field*, UCRL-5581-T; also *Phys. Rev. Letters* **3**, No. 9 (1959).
3. L. Rosen *et al.*, Los Alamos Scientific Lab., *Flux and Energy of Charged Particles at 300 and 600 Miles Altitude*, private communication.
4. J. A. van Allen *et al.*, Physics Dept., State Univ. of Iowa, *J. Geophs. Res.* **64**, 217 (1959).
5. J. A. van Allen, private communication.
6. L. Allen *et al.*, *Angular Distribution of Van Allen Radiation with Respect to Geomagnetic Field*, Post-deadline paper presented at *Amer. Phys. Soc. Meet.*, Washington (April 1959).

MEASUREMENTS OF FLUX OF SMALL EXTRA-TERRESTRIAL PARTICLES

HERBERT A. COHEN

Geophysics Research Directorate, Air Force Cambridge Research Center,
Air Research and Development Command

Abstract—Measurements from rockets and satellites using wire grids and microphones indicate a flux of 10^{-3} particles/(m² sec) of several microns in size into the Earth's atmosphere, with some indication of higher sporadic concentrations.

INTRODUCTION

In 1956 at the second meeting of the symposium an estimate[1] was made of the number of impacts of extra-terrestrial particles on an ICBM using three sources of information: (1) extrapolation of a relationship obtained from observing meteors, larger but less numerous luminescent particles, (2) interpretation of zodiacal and corona "*F*" light which give a mean spatial concentration and particle distribution for particles in the solar system, (3) one rocket measurement[2] using a technique similar to a scintillation counter. Since that time other direct measurements of small particle influx and damaging effects, have been made from satellites and rockets. Although the measured sampling area and time available is still small as compared to that available to meteor observers, an indication of the order of magnitude of the influx is now available from direct measurements using wire grids and sensitive microphones.

WIRE GRIDS

The wire grids, as shown in Fig. 1, developed by Manring[3] while at Air Force Cambridge Research Center, have made possible the first direct measurements of impact damage by micrometeors. In the latest use of the grids, on re-entry vehicles, the detector was a continuous length of 20 μ diameter, Evanohm wire, wound on a bakelite form, covering an area of 5·27 cm². The wire was put on in three layers so that no area of the card base was exposed. A 2·5–5·0 μ varnish layer was used on the top of the wires. If one wire of a detector was broken the detector became inoperative, with a subsequent measurable voltage change across a 7·5 Ω resistor in series with the detectors. The voltage change as

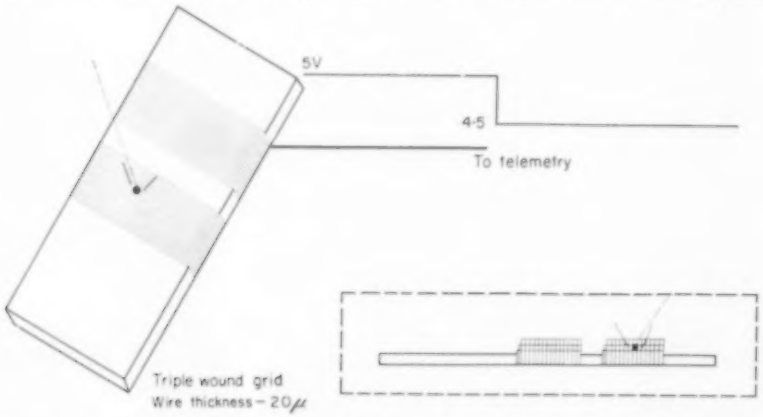


Fig. 1. Micrometeorite grid detector: An impacting micrometeorite larger than a minimum size breaks a strand of the continuously triply wound resistor. A voltage change is then transmitted to the telemetry.

a function of the number of grids broken is shown in Table 1.

If all eight detectors failed on a set and the electronic circuitry continued to operate the measurable voltage was changed to 1·5 V. If the electronic circuitry failed, the voltage dropped to 0·0 V.

Table 1. Voltage Output with Grid Detector Breaks

Number of grids broken	Voltage output (V)
0	4·95
1	4·75
2	4·3
3	4·1
4	3·6
5	3·3
6	2·3
7	1·92
8	1·5

Two sets of such grid impact detectors were exposed from opposite sides of the re-entry vehicle. Each set consisted of eight detectors connected in parallel.

From time of launch, $t = 0$ (t is the flight time), to $t = 778$ sec the two sets showed no broken grids. Between $t = 778{\cdot}651$ sec and $t = 778{\cdot}846$ sec the voltage from one set decreased to 1·5 V, the voltage from the other set remained constant. The vehicle then was at 263 km. No grids broke on the surviving set until re-entry; when the vehicle was at a height of 80·1 km all the remaining grids on the unbroken set probably broke. A few seconds later the measured voltage increased to over 5 V due probably to fusing of resistors during re-entry.

One of the two conclusions that can be reached from the results of this experiment are equipment failure, or an incident micrometeoric shower. There were two places where a rupture of a single wire would cause a decrease in voltage as seen on the records. The wire is No. 20 Teflon covered wire 0·032 in. in diameter soldered to junction posts, and protected from the outside environment by a bakelite strip. A small possibility exists that the wire became loose due to shock on ascent (heating is negligible) then fell off due to the roll of the vehicle. However, acceleration when the grids broke was negligible and the complete unit successfully survived an environment test before being placed on the vehicle. On a similar vehicle two sets of grids were protected for a full flight. The record showed that no wires broke during the entire flight.

Manring[3] has described the two previous experiments with grids on the satellites 1958α and 1958γ. On the 1958α, 10 cm^2 of 21 μ diameter grid wire (diameter with enamel insulation), was exposed for 60 days with at most only one of the grids being destroyed. On the 1958γ after 11 days of exposure two grids were destroyed on 1 day. The transmitters of the satellite ceased functioning within 5 days of the destruction of the grids. It was conjectured that the failure of the transmitter and grids were due to the passage of the satellite through a meteor shower.

The conclusions that can be drawn from the three flights of the exposed grids are (a) normally the mean influx of particles large enough to break the grid wires is most likely less than $1{\cdot}7 \times 10^{-3}$/(m^2 sec); (b) there are periods of showers with influx rates much greater than the normal mean, with one recorded instance of at least 8×10^4/(m^2 sec).

For the estimate of the destructive effects of micrometeors the length of time before the grids are destroyed is by itself valuable information. However, to compare the results from the method to other measuring techniques an estimate should be made of a common parameter such as the minimum size, momentum or energy necessary to break a wire. Two types of high velocity studies have been made: (1) At the Ballistic Research Laboratory, Aberdeen, Maryland, shaped charge tests were made under the supervision of Gehring.[4] From these tests it was concluded that a wire of the grid detector would be broken by particles of 50 μ diameter impacting at velocities between 10 and 14 km/sec. (2) At the Stanford Research Institute, experiments were performed under the supervision of Kells and Keough.[5] Three groups of particles were shot at a velocity of 4–6 km/sec at a 17 μ diameter wire grid with an 8 μ enamel

coating. The 15 μ diameter particles individually severed the wire while the 12 μ and 5 μ diameter particles individually did not sever the wires. The preliminary conclusion drawn then is that in the range 4–6 km/sec the minimum particle size required to break a wire is approximately one-half the wire diameter with coating.

MICROPHONES

A more widely used technique for measuring micrometeor influx is that of placing a sensitive crystal microphone directly in contact with some exposed surface of a rocket, or satellite, as shown in Fig. 2. The transducers used by AFCRC, designed and constructed by J. L. Bohn of Temple University, consist of a single brush crystal of cylindrical shape and mounted in a machined support such that the overall vibration resonances are tuned to an ultrasonic frequency range centered at 100 kc/s. The electrical pulse from the transducer is amplified through tuned stages of transistor circuitry. Vibrations due to particle impacts are transmitted along and through the vehicle surface to the crystal detector.* The signal is then converted to an electrical pulse, amplified, and transmitted to a ground station.

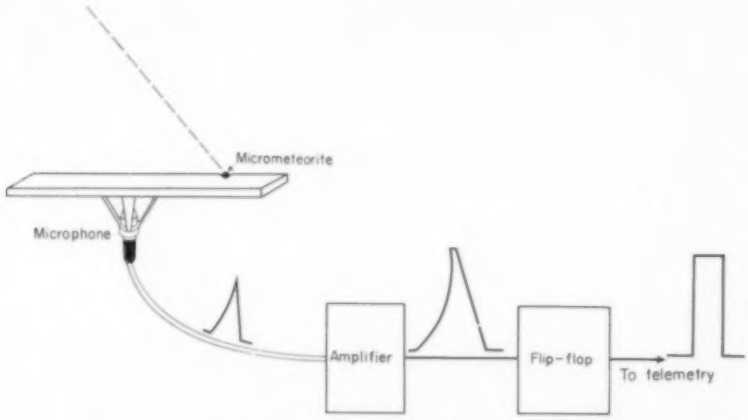


Fig. 2. Micrometeorite microphone detector unit: An impacting micrometeorite larger than a minimum size sets up an acoustic wave in a sensing plate. The signal is converted to an electrical pulse by a microphone, amplified and transmitted as a discrete signal change to the telemetry.

On the Pioneer I rocket launched in 1958 a crystal microphone was used with a two level amplifier. The crystal microphone was pressed against a surface of area 388 cm² acoustically isolated from the other sections of the vehicle. The system was calibrated by dropping or bouncing glass beads, ranging in diameter from 80 μ to 310 μ, on the outside sensitive surface from a height of 21·4 cm. By this technique it was found that the telemetry channel coupled to the high gain amplifier would respond to impulses greater than 5×10^{-4} g cm/sec, and both channels would respond to impulses greater than 5×10^{-3} g cm/sec. The flight time of the vehicle was $1{\cdot}56 \times 10^{5}$ sec; the vehicle attained a maximum altitude of 115×10^{3} km. During flight some portions of the data were lost due to poor signals; the total recorded clear signal time was $1{\cdot}14 \times 10^{5}$ sec. Over the total exposed area time of $4{\cdot}42 \times 10^{3}$ m² sec, a total of twenty-five impacts were recorded. One of these impacts was indicated by both channels of the amplifier, the remaining twenty-four were indicated only by the more sensitive channel. The average influx of particles with momentum greater than 5×10^{-4} g cm/sec but less than 5×10^{-3} g cm/sec was $5{\cdot}7 \times 10^{-3}$ particles/m² sec. There was one period, however, when the

* The microphone systems used by the Air Force Cambridge Research Center in their rocket and satellite measurements have been developed and constructed by members of the Physics Department of Temple University and members of the Research Foundation of Oklahoma State University, under contract with AFCRC.

influx was much higher than this mean; a total of seven impacts being recorded over a period of 6 minutes while the vehicle was rising through an altitude of approximately 109,000 km.

The microphone system, i.e. sounding plate, microphone and amplifier, have been calibrated using 80–100 μ diameter glass beads impacting at velocities between 5 cm/sec and 80 cm/sec. For these low impact velocities, as was expected from theoretical considerations, the response of the microphone system was linearly dependent on the momentum of the impacting particle. Experiments performed at Stanford Research Institute[5] with particles impacting at velocities between 1 and 2 km/sec have shown that at non-elastic impact the same functional relationship held as for elastic impact. However no theory has yet been established for extrapolating the relationship for lower velocities to cover impacts of greater than 11·2 km/sec, the minimum velocity of micrometeor entry into the Earth's atmosphere. If for a first estimate of particle size, it is assumed that the relationship for impacting particle momentum to microphone system output is the same for meteoritic velocities as it is for elastic impacts, further assumptions must be made concerning the particle density and velocity. The range of densities presently assumed for extra-terrestrial particles entering the Earth's atmosphere extends from 0·05 g/cm³, meteor puffball, to 7·9 g/cm, iron meteorite. The range of velocities for these particles is assumed to be between 11·2 km/sec, the escape velocity from the Earth to 73 km/sec, the sum of the escape velocity from the Sun at the Earth's distance plus the Earth's orbital velocity around the Sun. The minimum size of particles that can be measured by microphone systems developed for the Air Force Cambridge Research Center are listed in Table 2.

Table 2. Limiting Sensitivity of Microphone System

Microphone sensitivity (g cm/sec)	Minimum particle diameter measured: density = 0·05 g/cm³, velocity = 11·2 km/sec (μ)	Minimum particle diameter measured: density = 7·9 g/cm³, velocity = 71·9 km/sec (μ)
10^{-3}	32	3·2
10^{-4}	15	1·5
10^{-5}	7	0·7

The microphone was sensitive to impacts of momentum greater than $2{\cdot}5 \times 10^{-3}$ g cm/sec, when calibrated at elastic impact velocities. The Explorer I, launched 1 February, 1958, had a perigee altitude of 374 km, an apogee altitude of 2500 km and a period of 115 min.

Dubin[6] has reported an impact rate of $8{\cdot}4 \times 10^{-3}$ impacts/m² sec measured by use of a similar microphone on the Explorer I satellite; a total of 153 impacts were recorded over a period 7889×10^4 sec, on an area of 0·23 m².

LaGow *et al.*[7] reported an impact rate of $3{\cdot}54 \times 10^{-2}$ impacts/m² measured by a lead zirconate microphone on a sphere launched 27 May, 1958, that went to 2500 km altitude; a total of seventeen counts were recorded during 590 sec over an area of 0·814 m². Of these seventeen counts, seven occurred over a 6-sec period. The authors reported that the response of the system was proportional to the kinetic energy of the impacting particle and not the momentum. From measurements made with particles in the elastic impact range it was determined that the system could detect particles whose impacting energy was greater than 0·36 ergs.

A very wide area of disagreement of measured influx appears between the rates measured on the Pioneer rocket and the Explorer satellites,

and the values reported by Nazarova[8] of the Soviet Union. Nazarova reported that with an exposed area of 840 cm² on a satellite, an impact rate of 7·5 impacts/sec was measured, i.e. 190 impacts/m²/sec. It was also reported that on rockets using similar instrumentation, 31 impacts/m²sec was measured. The wide disagreement of influx rates is made even greater by the reported sensitivity of the Russian microphone system; the instrument could not measure impacts of momentum less than 0·1 g cm/sec, as compared to the limiting sensitivity of the microphone of approximately 10^{-3} g cm/sec as used on the Explorer and Pioneer vehicles. Unfortunately a description of the methods of calibration as employed by the Russians has not been made available.

put into the same form as definition (1) by the assumption of an average density and geocentric velocity for the observed particles. The minimum detectable particle probably has a mass of greater than 10^{-11} g.

If it is assumed that the distribution of particle masses impacting on Pioneer I and Explorer I follow the same form as definition (1) the values of c and α can be determined from the measurements: These values are shown in Table 3 with values deduced from meteor and zodiacal light observations.

It appears that if the relationship for the larger size meteors were extrapolated to micrometeorite size range that there would be good agreement between the mass distribution α values, from meteor, rocket and satellite observations. Assuming the same mass distribution for both sets of observations there would still be a difference of approximately one thousand in the flux values for one particular mass. Surprisingly there appears to be poorer agreement between the impact observations and the zodiacal light observations of similar size particles, than between the impact observations of small particles and extrapolation from observations of larger particles, meteors.

Table 3. Mass Distribution of Micrometeorite Influx* $N = 10^{-c}m^{-\alpha}$

Source	c	α	Assumptions	
			particle density g/cm³	particle velocity km/sec
Meteor[9]	12·11	1·34		
Zodiacal light[10]	11·84	0·33	3·0	15
Pioneer I	15·55	1·40		
Explorer I[9, 6]		0·79	3·0	15
Explorer I[6]	14·34	1·40 (assumed)	3·0	

* N is the influx/m²sec of all particles of mass m g or larger.

COMPARISON OF METEOR ZODIACAL LIGHT AND IMPACT MEASUREMENTS

It has been established[9] that the rate N at which meteors of mass m or larger enter the Earth's atmosphere follows the relationship

$$N = 10^{-c}m^{-\alpha} \quad (1)$$

From the paper of Hawkins and Upton[9] it is assumed that the minimum detectable meteor has a mass greater than 10^{-4} g.

From observations of the corona "F" light and the zodiacal light a relationship of the type

$$n = 10^{-d}r^{-\beta} \quad (2)$$

has been found[10] for the mean concentration n of particles of radius r or larger. This can be

CONCLUSIONS

Measurements from American rockets and satellites indicate that 10^{-3} particles/m² sec is the order of magnitude of flux into the Earth's atmosphere of extra-terrestrial particles several microns in size. There are also indications of

shower periods. The mass distribution of the particles appears to be similar to the mass distribution observed for meteors but is in poor agreement with the mass distribution for particles in interplanetary space as determined from zodiacal light measurements.

REFERENCES

1. H. A. COHEN, *Transactions of Second Technical Symposium on Ballistic Missiles* **4**, 109 (1957), Ramo-Wooldridge, Los Angeles.
2. O. BERG and L. J. MEREDITH, *J. Geophys. Res.* **61**, 751 (1956).
3. E. R. MANRING, *Planet. Space Sci.* **1**, 27 (1959).
4. H. A. COHEN, A. CORMAN and M. DUBIN, *Proc. 3rd Symp. on Hypervelocity Impact*, **1**, 405 (1959), Armour Research Foundation of Illinois Institute of Technology, Chicago.
5. M. C. KELLS and D. D. KEOUGH, *Impact of High Speed Particles with Solids and Gases*, Air Force Cambridge Research Center Contract TR 59-219, ASTIA Document No. AD-210-935.
6. M. DUBIN, *Meteoritic Dust Measured From Explorer I*, to be published.
7. H. E. LAGOW, D. H. SCHAEFER and J. C. SCHAFFERT, *Micrometeorite Impact Measurements on a 20 in. Diameter Sphere at 700 to 2500 Kilometers Altitude*, paper presented at the Rocket and Satellite Symposium, 5th Reunion of the *Comité Special Année Géophysique Internationale*, Moscow, July–August 1958. (Papers reproduced by the National Academy of Sciences).
8. I. N. NAZAROVA, *Rockets and Satellite Investigation of Meteors, Summary*, Paper presented at the Rocket and Satellite Symposium, 5th Reunion of the *Comité Special Année Géophysique Internationale*, Moscow, July–August 1958. (Papers reproduced by the National Academy of Sciences).
9. G. S. HAWKINS and E. K. L. UPTON, *Astrophys. J.* **128**, 727 (1958).
10. H. SIEDENTOPF, *Meteors*, pp. 145, 146 (ed. by T. R. KAISER), Pergamon Press, London (1955).

DEDUCTION OF VERTICAL THERMAL STRUCTURE OF A PLANETARY ATMOSPHERE FROM A SATELLITE*

JEAN I. F. KING
Missile and Space Vehicle Department, General Electric Company

Abstract—An inversion technique is developed for deducing the temperature stratification of a planetary atmosphere from the variation of the emission spectrum with air mass. This is a planetary application of the well-known solar limb-darkening effect. Mathematically, the emergent intensity is an integral transform of the Planckian temperature taken as a function of optical depth. The solution therefore consists in devising a scheme for inverting the transform.

Pending satellite observations, the following approach is used for the verification of the theory. Ground observations of the emission of the 9·6 μ ozone band as a function of air mass thickness have been taken. These radiative data provide information on ozone "at a distance", in a manner similar to the remote satellite viewing from above. Four sets of ozone emission data have been analysed and the vertical thermal structure of the ozonosphere deduced therefrom. A critique of the method together with future research directions is given.

I. INTRODUCTION

It is common knowledge that the solar disk is not uniformly bright. This "limb-darkening" effect arises from the temperature stratification in the photosphere, which cools with increasing radius. The intensity variation from the center to edge of the solar disk can be made to yield the thermal gradient of the photosphere.[1]

In principle this technique can be used to probe thermally the emission spectrum of a planetary atmosphere or, more generally, the radiation field of any stratified emitter. We have therefore the possibility of continuously monitoring the vertical thermal structure of an atmosphere by means of limb-darkening scans taken by an orbiting space vehicle.

In order to learn the limitations of the method, an attempt is made herein to deduce the thermal structure of the atmospheric ozone layer solely from ground observations. These consist of radiation intensity measurements at the center of the 9·6 μ ozone emission band viewed at varying zenith angles. Our experimental data, given in Table 1, are careful measurements made by Adel[2] at the Atmospheric Research Observatory, Flagstaff, Arizona.

* This work has been supported by the Air Force Ballistic Missile Division under Contract No. AF 04 (647)-269.

II. THEORY AND APPROXIMATIONS

The radiation intensity emitted by a non-scattering atmosphere is an integral transform of the Planck temperature field. This is a consequence of the fundamental equation of radiative transfer and the assumption of local thermodynamic equilibrium.

For a horizontally stratified atmosphere, i.e. one in which properties vary in the vertical only, the monochromatic emergent intensity is simply the Laplace transform of the Planck function considered as a function of optical depth. Symbolically,

$$I_\nu(0, \mu) = \int_0^\infty B_\nu(\tau) \exp(-\tau_\nu/\mu)\, \mathrm{d}\tau_\nu/\mu = (1/\mu)\, \mathscr{L}_{1/\mu}[B_\nu(\tau)]n \qquad (1)$$

where $I_\nu(0, \mu)$ is the monochromatic radiant intensity emerging from the atmosphere at an angle $\theta = \arccos \mu$ with the outgoing normal and

$$B_\nu(\tau) = \frac{2hc^2\nu^3}{\exp[h\nu/kT(\tau)] - 1}$$

is the Planck intensity, a function of wave number and, implicitly through the optical depth, of temperature. The intensity variation with angle is an indication of the departure of the atmosphere from an isothermal state. The

monochromatic optical thickness is defined as usual by

$$\mathrm{d}\,\tau_\nu = \kappa_\nu \rho \mathrm{d}u$$

where κ_ν and $\mathrm{d}u$ are the mass absorption coefficient and differential path length.

For the ozonosphere problem, we reckon the optical depth as positive upward, with the ground surface taken as $\tau = 0$. The Earth satellite or planetary observation would be the reverse of this. In that event the optical depth would be zero at satellite height and increase in the nadir direction.

In this approximate treatment we assume a gray (κ_ν independent of frequency), homogeneous (κ_ν independent of position) atmospheric model. Thus we may drop the subscripts in equation (1).

We note, therefore, that the vertical temperature pattern can be obtained in principle from a knowledge of intensity variation with angle by inverting the transform. Thus from equation (1)

$$B(\tau) = \mathscr{L}^{-1}_{1/\mu}[\mu\, I(0, \mu)] \tag{2}$$

The words, in principle, should be stressed, for the inversion of any but a few simple functions is generally difficult and inexact. For inasmuch as a transform is an integration operation, in taking the inversion we are in a sense differentiating. This tends to multiply errors.

III. RESULTS

A number of methods have been tried and discarded for the inversion operation indicated by equation (2).[3] The most fruitful technique has used the Lagrangian interpolation formula to fit exactly a fourth degree polynomial to the five experimental points given in each of the soundings of Table 1. In this manner the polynomial coefficients a_i were determined for each day's observation, viz.

$$I(0, \mu) = a_0 + a_1 \mu + a_2 \mu^2 + a_3 \mu^3 + a_4 \mu^4 \tag{3}$$

The inversion of this transform, an elementary one, yields at once

$$B(\tau) = a_0 + a_1 \tau + \frac{a_2 \tau^2}{2!} + \frac{a_3 \tau^3}{3!} + \frac{a_4 \tau^4}{4!}\,. \tag{4}$$

Table 1. Ozone Emissivity* at 9·6 μ vs. Air Mass Thickness

Thickness (sec $\theta = 1/\mu$)	3 Oct. 56	15 Dec. 56	15 Jan. 57	10 Apr. 57
5·8	21·12	19·12	20·50	25·01
3·8	15·94	16·96	17·07	20·39
2·9	13·84	14·16	15·79	17·70
2·0	12·81	11·54	11·60	14·10
1·0	8·73	8·00	9·10	10·14

* ergs cm^{-2} sec^{-1} wave-no.$^{-1}$

From these equations we see what is intuitively obvious, that the ground temperature intensity is equal to the emergent intensity viewed along the horizon, i.e.

$$B(0) = I(0, 0) \tag{5}$$

We have further, with the gray atmosphere case, that the rate of change of intensity near the horizon is proportional to the lapse rate near the ground:

$$\left.\frac{\mathrm{d}I(0, \mu)}{\mathrm{d}\mu}\right|_{\mu \to 0} = \left.\frac{\mathrm{d}B}{\mathrm{d}\tau}\right|_{\tau \to 0}. \tag{6}$$

A subtle error arises from the semi-infinite character of the atmosphere assumed in equation (1). The inversion, equation (4) diverges positively or negatively for sufficiently large optical depth. The actual atmosphere is bounded, however, which is equivalent to the artifice of assuming $B(\tau) = 0$ for τ greater than τ_0, the thickness of the atmosphere. This defect can be

corrected only by a formalism explicitly taking the finiteness of the atmosphere into account.

Figures 1–4 are double plots of both the emissivities (plotted against μ) and their inversions, the Planck intensities plotted as a function of optical depth. Since the black-body function varies monotonically with temperature, following an approximately T^5 dependence in this

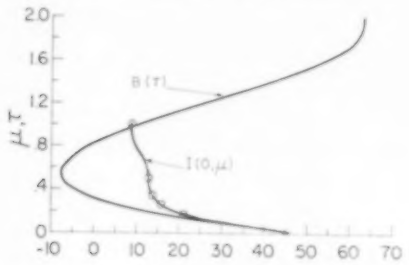


Fig. 1. Vertical thermal pattern of the ozone layer over Flagstaff, Arizona, 3 Oct. 1956.

spectral region, we can gain from these curves information on the temperature distribution with depth in the ozone region.

Figures 2 and 3 can be omitted from further consideration since both show a spurious hump

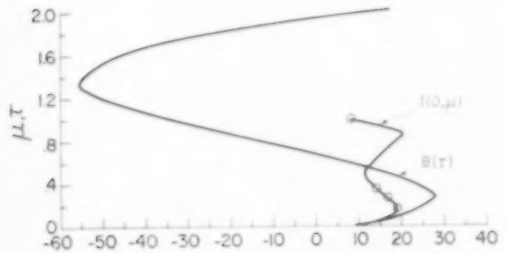


Fig. 2. Vertical thermal pattern of the ozone layer over Flagstaff, Arizona, 15 Dec. 1956.

in the Lagrangian polynomial. These unphysical aberrations would disappear if there were more points to fit. They arise from the fact that there are no observations between 0° and 60° from zenith.

On the other hand, Figs. 1 and 4 have smoother polynomial fits to the observations. The shapes of the curves are seen to be "meteorological". One notes the rapid temperature decrease in the thin tropopause height. Thence an increase ensues in the stratosphere, followed eventually by a temperature decline to outer space.

Quite disconcertingly, we note that Fig. 1 predicts a "negative" tropopause temperature while Fig. 4 yields one much too low. These

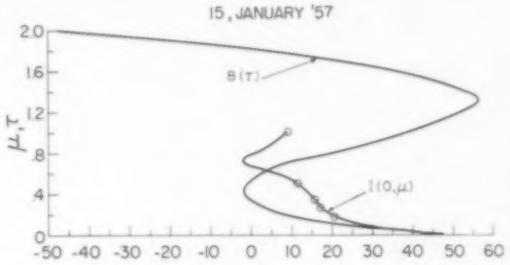


Fig. 3. Vertical thermal pattern of the ozone layer over Flagstaff, Arizona, 15 January 1957.

relate to the fact that the exponential absorption law, implicit in the gray atmosphere assumption, gives far too rapid attenuation.

IV. CRITIQUE AND CONCLUSIONS

We now list the assumptions made and discuss the possibility of their relaxation:

(1) Local thermodynamic equilibrium (non-scattering atmosphere). According to Curtis and Goody[4] this assumption is valid in the atmosphere below 75 km where collisions permit radiative relaxation of the vibrational bands.

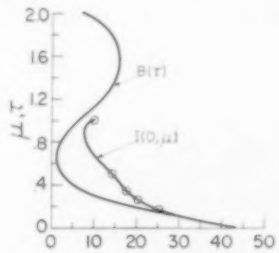


Fig. 4. Vertical thermal pattern of the ozone layer over Flagstaff, Arizona, 10 April 1957.

Above this level the coupling between the temperature and radiation field disappears, and our formalism becomes inapplicable.

(2) Neglect of Earth's curvature. For the ozonosphere this is a small correction term.[5]

The effect becomes large for satellites hundreds of kilometers above a planetary surface, but can easily be taken into account.

(3) Plane-parallel atmosphere. The neglect of temperature and ozone variations in the horizontal (or more correctly, along a radial shell) is a serious one. Justification exists only if the area scanned is small compared to that of the planetary disk, a condition obviously not met in an orbiting vehicle. These horizontal inhomogeneities can be partially averaged out by having the detector telescope rotate in a manner such that the loci of the barrels form conic surfaces of varying apexes. For a satellite stabilized about a vertical axis, the rays would trace out concentric circles, with the nadir as centre

(4) Gray atmosphere. For the line absorption spectra of far-infrared bands this assumption is palpably false. As seen above, the "negative" temperatures are probably traceable to this cause.

(5) Homogeneous atmosphere. The monochromatic or any generalized absorption coefficient are functions of pressure and temperature (and hence of position) as well as the molecular species. To the first approximation, these variations can be scaled into a "reduced" path length.

(6) Infinite optical thickness. This implicit assumption led to the unreal polynomial divergences referred to above.

We turn now to the observations themselves which were gratuitous, and not taken with this probing application in mind. For our purposes, desiderata would be about twice as many emissivity measurements at equally spaced increments of cos θ, each one preferably scanned circularly and averaged about a constant zenith angle. This would eliminate the spurious humps by a higher degree polynomial fit, give equal weight to all zenith angles and tend to iron out horizontal inhomogeneities.

We conclude that the method bears promise and warrants further investigation. This is attested by the fact that even with gross assumptions, profiles are deduced which reproduce the main features of the ozonosphere thermal structures. Any treatment which aspires to completeness, however, must include effects of atmospheric band structure, inhomogeneity and finite thickness directly in the formalism.

Although we have used the inversion technique in connection with scans over zenith angle, it is equally applicable to scans over frequency as well. The possibility of probing band regions of varying opacity has been considered by Kaplan [6].

We reserve the treatment of non-grayness and finite atmosphere effects in limb-darkening and frequency scans to a future paper.

Acknowledgment—It is a pleasure to acknowledge the computing help of Miss Janice M. Furey.

REFERENCES

1. V. Ambartsumyan, *Theoretical Astrophysics*, Pergamon Press, London (1958).
2. A. Adel, private communication (1959).
3. J. I. F. King, Chap. 14 in *Scientific Uses of Earth Satellites* (ed. by J. A. van Allen), Univ. of Michigan Press (1956).
4. A. R. Curtis and R. M. Goody, *Proc. Roy. Soc.* **A236**, 193 (1956).
5. E. S. Epstein, C. Osterberg and A. Adel, *Manual for VODARO, Vertical Ozone Distributions from the Absorption and Radiation by Ozone*, Sci. Rep. HA-7, AFCRC Report, Contract No. 19 (122)-198 GRD, Arizona State College (1955).
6. L. D. Kaplan, *J. Opt. Soc. Amer.* **49**, 1004 (1959).

THERMAL IMPACT—THE MECHANICAL RESPONSE OF SOLIDS TO EXTREME ELECTROMAGNETIC RADIATION

J. E. MICHAELS

Structures and Dynamics Operation, Aerosciences Laboratory Missile and Space Vehicle Department, General Electric Company

Abstract—The phenomenon of thermally induced stress waves resulting from the deposition by radiation of intense electromagnetic energy into the surface of a solid is described. Both theoretical and experimental aspects are treated. The derivation and general solution for the case of a semi-infinite bar are given. A recent experiment is described which for the first time demonstrated and verified the thermal impact phenomenon. The design and development of an optimized experimental facility by which the phenomenon can be studied in the laboratory is discussed.

I. INTRODUCTION

In connection with the A.I.C.B.M. problem are studies concerned with the kill-effectiveness of a thermo nuclear explosion in space. This problem is being pursued with great vigor from the standpoint of determining the characteristics of the vacuum explosion and of observing the mode of failure of materials subjected to intense stress (or shock) waves. More recently, interest has grown toward obtaining a better understanding of the actual mechanism by which stress waves are induced in a solid, subjected only to a radiant heating pulse of great magnitude, and to the development of an experimental facility which can reproduce the phenomenon in the laboratory.

If the debris of explosion products is not considered, the energy transferred to an intercepting object from the vacuum explosion is entirely in the form of electromagnetic waves. This energy is absorbed into the surface and converted into sensible heat. If the object is sufficiently close to the explosion, a thin surface layer of material is instantly vaporized and accelerated away. The resulting reactive force produces a stress wave which then propagates through the solid causing catastrophic damage by spallation fracture. For a more distant explosion, the heated surface may not vaporize or even melt, but due to its sudden expansion, may still induce a stress wave of damaging magnitude in the solid. This paper is concerned with this latter case and treats both the theoretical and experimental aspects of this phenomenon which is called thermal impact.

II. THEORETICAL ANALYSIS

Several investigators[1–4] have examined the transient stresses resulting from suddenly heating the surface of a solid. However, solutions have usually been restricted to the maximum heating rates that could be expected by ordinary conduction and convection of heat to the sur face. For the radiant transfer of energy, heating rates of much greater magnitudes can occur. Since shorter wave lengths of radiant energy are usually associated with greater intensities and can penetrate into the surface somewhat before being absorbed, the actual temperature response of the solid is of utmost importance. An analysis which permits complete flexibility in this respect was performed by the writer,[5] and is briefly summarized here.

The ordinary wave equation written in one dimension with the thermo-elastic term as the driving function is

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial \tau^2} = \frac{\partial^2 u}{\partial x^2} - \alpha \frac{\partial T(x,\tau)}{\partial x} \tag{1}$$

The temperature $T(x,\tau)$ is a completely arbitrary function of time and x-space. Thus, any law of energy transfer into the solid may be utilized.

Equation (1) is applicable to longitudinal waves in a slender bar only if the effective wave

lengths are long in comparison to the transverse dimensions of the bar. Then the effect of lateral inertia and shear may be neglected. Since this depends upon the temperature distribution function involved, a particular result must be examined for its applicability to this analysis.

Consider a semi-infinite bar extending from $x = 0$ to $x = \infty$. The bar is assumed to be disturbed from rest in its equilibrium position with the end at $x = 0$ stress free. The following initial and boundary conditions prevail:

Initial conditions

$$u]_{\tau=0} = 0 \tag{2}$$

$$\left.\frac{\partial u}{\partial \tau}\right]_{\tau=0} = 0 \tag{3}$$

Boundary conditions

$$\sigma]_{x=0} = E\left[\frac{\partial u}{\partial x} - \alpha T(x,\tau)\right]_{x=0} = 0$$

$$\therefore \quad \left.\frac{\partial u}{\partial x}\right]_{x=0} = \alpha T(0,\tau) \tag{4}$$

Using the method of integral transforms the general solution to equation (1) subject to the conditions of (2), (3) and (4) may be shown to be

$$u(x,\tau) = -\frac{2c\alpha}{\pi}\int_0^\infty \cos(\xi x)\int_0^\tau \sin\left[c\xi(\tau-\eta)\right]\int_0^\infty T(x,\eta)\sin(\xi x)\,dx\,d\eta\,d\xi \tag{5}$$

As an example, consider the temperature function

$$T(x,\tau) = a\tau \exp(-bx) \tag{6}$$

where a and b are constants. The function was chosen to allow for some law of energy y penetration into the solid as well as the ordinary law of heat diffusion thru the solid. The solution to equation (5) with $T(x,\tau)$ defined by equation (6) is

$$u = -\frac{a\alpha\tau}{b}\exp(-bx) + \frac{a\alpha}{b^2c}\begin{cases} -\exp(-bc\tau)\cosh(bx) + 1 & x < c\tau \\ \exp(-bx)\sinh(bc\tau) & x > c\tau \end{cases} \tag{7}$$

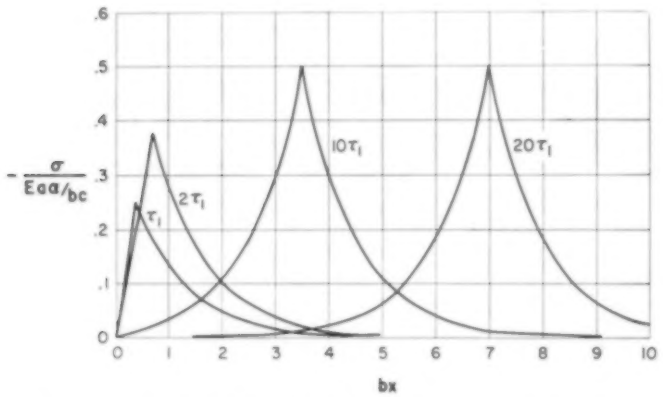


Fig. 1. Propagation of thermally induced stress wave of example.

from which the stress is obtained as

$$\sigma = -\frac{Ea\alpha}{bc}\begin{cases} \exp(-bc\tau)\sinh(bx) & x < c\tau \\ \exp(-bx)\sinh(bc\tau) & x > c\tau \end{cases} \tag{8}$$

Figure 1 is a plot of $-\sigma/(Ea\alpha/bc)$ versus bx which shows the shape of the stress wave as it is induced and propagated along the bar. The maximum stress occurs when $x = c\tau$ and approaches a limiting value as $\tau \to \infty$.

A relation between heating rates and the magnitude of induced stresses obtained for this example show that about 5000 W/cm² will induce a stress of 1 psi in aluminum. This indicates that heating rates of the order of 10^8 W/cm² may induce damaging stresses in aluminum.

In this example the heating rate is the most important parameter; not the duration of the pulse nor the total energy deposited. One must be cautious in interpreting these results, however, since the magnitude of induced stresses is undoubtedly very sensitive to the actual temperature response of the bar.

III. FIRST EXPERIMENTAL RESULTS

In order to observe the thermal impact phenomenon a simple laboratory experiment was devised. It consisted essentially of a light source of high intensity, a lens system to gather and focus the light, and a $\frac{1}{4}$ in. diameter aluminum rod about 3 ft long. A schematic drawing of this experiment is shown in Fig. 2.

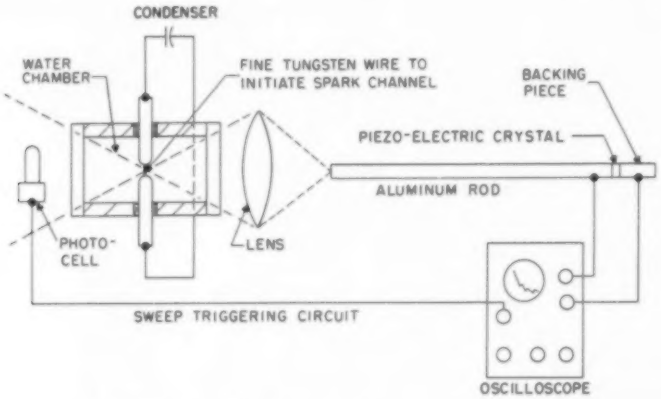


Fig. 2. Schematic of thermal impact experiment.

The light source utilized an underwater exploding wire spark to produce a radiant pulse which was focused on one end of the aluminum rod. At the other end of the rod a piezo-electric crystal was attached with its output connected to an oscilloscope equipped with a recording camera. The oscilloscope sweep was triggered by a photo-cell actuated by stray light from the spark. The intensity of the radiant flux hitting the rod was measured by a photo-cell showing a 50 μsec pulse with a maximum flux of about 600 W/cm².

The exploding wire spark was powered by the discharge of a capacitor of 5 μF capacitance charged to a potential of 20 kV. To produce a spark of great intensity, the resistance of the spark gap must dominate the discharge circuit. The resistance of an ordinary spark in air drops to a very low value once the conducting path or spark has been initiated. Maintaining a high resistance during the complete discharge of the stored energy is most easily accomplished by confinement of the spark path; hence the under water spark which confines the conducting plasma by virtue of the inertia of the surrounding water.

Figure 3 shows two oscilloscope records which were obtained from this experiment. The

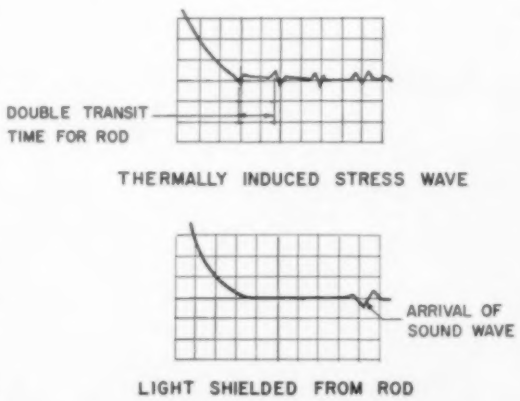


Fig. 3. Oscilloscope records of thermally induced stress waves. (Horizontal sweep 200 μsec per division.)

main sweep was influenced by an electrical bias picked up from the spark by inadequately shielded instrumentation. Since a current surge of thousands of amperes occurs during the

discharge cycle, ordinary shielding methods are insufficient and considerable effort and cost are required to eliminate this problem. However, the induced stress waves merely appeared superimposed on this bias such that no improvement in shielding was deemed necessary for the purpose of this experiment.

The upper record of Fig. 3 shows the induced stress waves superimposed on the aforementioned bias. The time between waves is that required for a wave to be reflected to the front end of the rod and back again to the crystal. These waves represent a stress magnitude of about 0·01 psi.

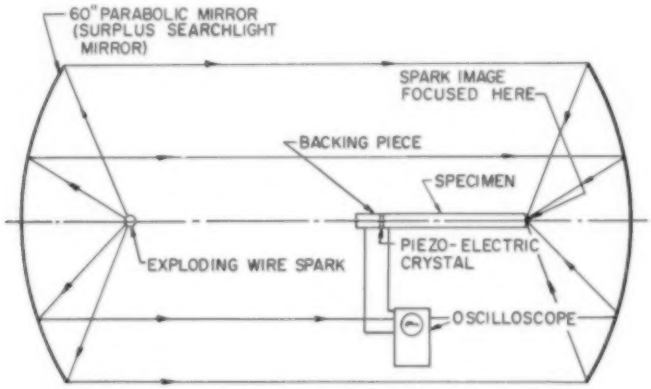


Fig. 4. Diagram of thermal impact facility.

The lower record shows the results of an identical test except that the light was shielded from the end of the rod. The stresses finally observed in this test were those initiated by the very audible sound waves of the external switch of the discharge spark circuit and their arrival time corresponds to the distance that the switch was located with respect to the end of the rod.

No attempt was made to optimize this experiment. Its purpose was merely to demonstrate the thermal impact phenomenon in the laboratory. To the writer's knowledge these are the first experimental results that have been obtained which demonstrate the existence of thermally induced stress waves.

IV. THE THERMAL IMPACT FACILITY

The rather simple experiment described above not only verifies the thermal impact phenomenon but it also suggests an excellent means by which it can be studied in the laboratory; that of using the intense radiant energy from an exploding wire spark. A comparison between experiment and theory of the heating rates required for unit stress indicates about an order of magnitude difference. Since the temperature distribution function at the critical rod end for the experiment was not determined, this result is not too surprising. Indeed, one could be quite satisfied with this correlation. The obvious conclusion is that an experimental facility designed to study the thermal impact phenomenon must be able to deliver electromagnetic energy to the surface of a solid at an exceedingly high rate. Furthermore, since stress waves are the resultant effect which is observed, the energy must be deposited with no appreciable mechanical forces which would interfere with the thermally induced waves.

The design of an optimized thermal impact facility was patterned after the experiment described previously. A schematic diagram is shown in Fig. 4. It consists essentially of the same elements of the simple experiment, an exploding wire energy source and a mirror system to collect and focus the energy on the end of a bar specimen.

To power the exploding wire spark, twelve low inductance capacitors, each rated at one microfarad capacitance are used. The capacitors are divided into two banks of six in parallel,

with the banks connected in series. When charged to 80 kV d.c. (total across both banks) the total energy stored is 9600 J. The design inductance of the total circuit is below 10^{-8} H which gives a ringing frequency of about 1 Mc/s. If the circuit were critically damped (which, of course, depends upon the resistance of the exploding wire load), it would dissipate all of the stored energy in about one microsecond. The maximum power (rate of energy dissipation) for this case would be about 10^{10} W.

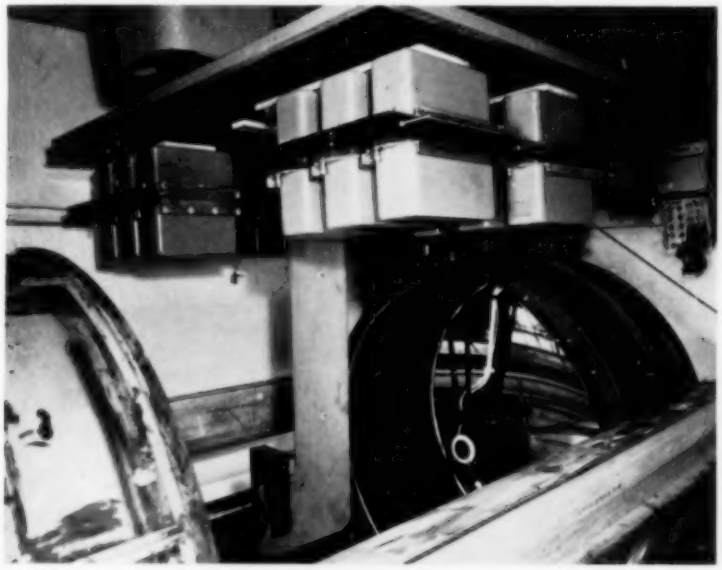


Fig. 5. Capacitor bank and copper conductor plates.

The extreme low circuit inductance is obtained through the use of large flat copper plates for the electrical conductors and a switch and exploding wire load of coaxial design. Fig. 5 shows the general arrangement of the capacitor banks and the copper conductor plates. The plates are coated with polyethylene (as an efficient dielectric) and clamped together to minimize inductance.

The switch, which is not visible but is mounted directly below the upright section shown, is of the vacuum type and consists of two tungsten electrodes about 1 in. in diameter. An electrical solenoid actuates the switch by plunging the electrodes close together and allowing a spark to jump the small remaining gap with a minimum of resistance to the total circuit.

The exploding wire load is contained in a cartridge mounted at the top of the plates which neck upward from the capacitor bank for positioning at the focal point of the 60 in. parabolic collector mirror. A close-up view of this cartridge is shown in Fig. 6. The cartridge is open on the side facing the mirror such that the full 120° acceptance angle of the mirror is utilized. A fine tungsten wire, 2 mils in diameter and $\frac{1}{4}$ in. long, connects between the electrodes and initiates the main spark path. Proper positioning of the spark at the exact focal point of the collector mirror is made possible by mounting the entire capacitor bank and discharge circuit assembly on a milling machine table which is adjustable in three dimensions.

Figure 7 shows a general view of the facility in which the collector mirror (on the right) is set back to show the capacitor bank and table. Visible on the extreme right of the picture is the power supply used to charge the capacitor bank. The parabolic mirrors are mounted on overhead beams to permit easy alignment and access beneath for the capacitor bank and table

assembly. In normal position, the collector mirror is directly over the capacitor bank with the necked portion of the conductor plates fitting thru the slot cut in the rim of the mirror frame. In this position the exploding wire spark is at the focal point of the collector mirror.

Fig. 6. Close-up view of exploding wire cartridge.

At the focal point of the other mirror is mounted the instrumented specimen (not shown in the pictures). Since this mirror arrangement gives a 1 : 1 image ratio, a $\frac{1}{4}$ in. diameter specimen (equal to the spark length) will be used. A cylindrical shroud connecting the two mirrors will completely enclose the optic volume, which will be purged of air and filled with helium at slightly greater than atmospheric pressure. This will permit passage of the shorter wavelengths which would normally be absorbed by the oxygen in air.

Much theoretical and experimental work concerning the exploding wire phenomenon (EWP) is available in the literature.[6] It is not the purpose of this paper, however, to delve into this very interesting subject except to indicate that the attainment of ultra-high temperatures by EWP are quite feasible. The design of the thermal impact facility is based on this concept.

In estimating the probable spark intensity, the area of the radiating surface of the spark is

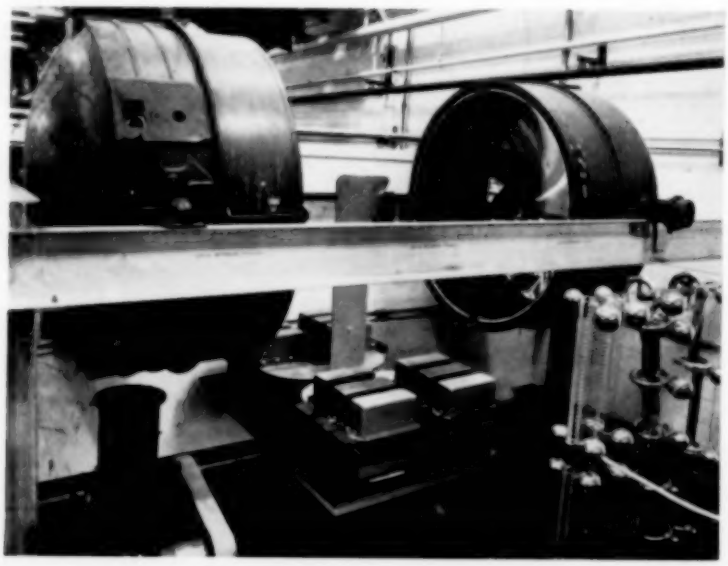


Fig. 7. General view of thermal impact facility.

required. Assuming that the spark plasma is about 0·05 in. in diameter (at the time of its maximum temperature) gives a radiating area of 0·5 cm². For a maximum rate of energy dissipation in the spark equal to 10^{10} W, this gives an emissivity of 2×10^{10} W/cm². Further assuming that half of this power is lost in mechanical work (producing a shock wave), this gives a free radiating intensity of the spark of 10^{10} W/cm² which corresponds to the radiation of a black body at approximately 200,000°K. Temperatures of this magnitude and higher by EWP have been reported in the literature.[7]

Optic losses from two main sources reduce the radiation intensity actually delivered to the specimen. First, the 120° acceptance angle of the mirror intercepts only 25 per cent of the total spherical radiation. Secondly, the reflectivity of the aluminum surfaced mirrors and the absorptivity of the helium atmosphere will only permit passage of radiant wave lengths greater than 0·1–0·2 μ. The resulting gross loss amounts to between 3 and 4 orders of magnitude. This means that the collected intensity of radiation which arrives at the specimen will be between 10^6 and 10^7 W/cm².

V. CONCLUSIONS

No other laboratory facility is known which can produce heating rates of this magnitude on a specimen in the absence of appreciable mechanical forces. Furthermore, at the present time, the magnitude of the resulting induced stresses for such energy fluxes cannot be predicted. It is believed that a theoretical base has been established for the thermal impact phenomenon and the experimental facility described will permit its further study in the laboratory.

Nomenclature

- a constant (°F/sec)
- b constant (1/in.)
- c stress wave velocity (in/sec) $= \sqrt{(E/\rho)}$
- E elastic modulus (psi)
- T temperature (°F)
- u longitudinal displacement (in.)
- x longitudinal co-ordinate of bar (in.)
- α coefficient of expansion (in./in.-°F)
- η dummy variable of integration
- ξ transformation parameter
- ρ mass density of bar (lb-sec²/in⁴)
- σ longitudinal stress (psi)
- τ time (sec)

REFERENCES

1. T. Mura, *Thermal Strain and Stresses in Transient State*, reported at *2nd Jap. Nat. Congr. Appl. Mechanics* (1952). (In English).
2. J. Ignaczak, *Thermal Displacements in an Elastic Semi-Space Due to Sudden Heating of the Boundary Plane*, *Arkh. Mech. Stosowanej* **9**, 395 (1957). (In English).
3. V. I. Danilovskaya, *Thermal Stresses in an Elastic Half-Space Arising After a Sudden Heating of its Boundary*, *Appl. Math. Mech., Leningr.* **14**, 316 (1950). (In Russian).
4. V. I. Danilovskaya, *On a Dynamical Problem in Thermoelasticity*, *Appl. Math. Mech., Leningr.* **16**, 341 (1952). (In Russian).
5. J. E. Michaels, *Thermally Induced Elastic Wave Propagation in Slender Bars*, p. 209 in *Proc. 3rd U.S. Nat. Congr. Appl. Mechanics*, Amer. Soc. Mech., Engrs. New York (1958).
6. W. C. Chace, *A Bibliography of the Electrically Exploded Wire Phenomenon*, Geophys. Res. Direct. Res. Notes No. 2, AFCRC-TN-58-457 (Nov. 1958), Astia Doc. No. AD-152640.
7. H. Fischer, *Upper Temperature Limits in the High Pressure Discharge*, presented at *Conference on Extremely High Temperatures*, 18–19 March 1958, Boston, Mass., Wiley, New York (1958).

A DYNAMICAL MODEL FOR INVESTIGATION OF THE DOWNWARD TRANSFER OF ENERGY INTRODUCED BY UPPER ATMOSPHERIC HEATING

LOUIS BERKOFSKY and RALPH SHAPIRO
Geophysics Research Directorate, Air Force Cambridge Research Laboratories

Abstract—A dynamical model for the investigation of the effects of heating at high atmospheric levels on lower level flow is devised. The source of energy is assumed to be anomalous corpuscular radiation which is associated with increased auroral activity, and introduces increased mechanical and thermal energy in the auroral zone. The governing equations form a complete set of mutually consistent statements concerning the thermal and flow fields and their interactions.

I. INTRODUCTION

The search for evidence of the existence of significant meteorological responses to variable solar-energy emission is being actively pursued today and has been for decades. The subject of solar-weather relationships has attracted the attention of scientists and others with widely differing backgrounds who have brought to the problem varied approaches. The American Meteorological Society in a recent bibliography,[1] which it admits is incomplete, lists almost 1300 references on the subject of solar-weather relationships.

In spite of the longevity of the problem and the attention it has been given, the central issue of the existence of such effects remains in doubt. Much of this doubt persists because favorable evidence (such as it is) is of a statistical nature. In such problems without a clear physical theory, it is extremely difficult to assess the statistical significance of the results in a completely unambiguous fashion.

The suggestion has been made that a theoretical study, proceeding from basic physical principles and based upon a model which makes use of recent knowledge of atmospheric dynamics and of solar energy variations, might prove to be definitive in establishing not only the existence of solar effects on the weather but also in establishing the physical processes by which such effects are produced. This study attempts to carry out the above suggestion within the framework of an atmospheric model which is sufficiently realistic to accomplish its purposes but no more refined than necessary. This model can be considered to be the first of a series of related experiments. The first experiment has been designed for its simplicity, consistent with current knowledge of solar energy variations and the physical characteristics and behavior of the very high atmosphere.

This model is designed to test a recent statistical result[2, 3] which presented substantial evidence of a response of the surface-pressure distribution following sudden, large increases of geomagnetic activity. The solar perturbation in this case is assumed to be a large increase in corpuscular radiation which would have a direct effect on the atmosphere primarily at high levels in the auroral zone. The actual energy input is felt to be too small to have any measurable direct effect at lower atmospheric levels. The model attempts to reproduce this situation by maintaining the energy input within reasonable bounds. Consequently, any response by lower atmospheric levels must be accomplished by means of a redistribution of existing atmospheric potential or kinetic energy.

II. THE MODEL

A. Basic Equations

The model atmosphere is assumed to be in hydrostatic equilibrium and frictionless. The motions are ageostrophic; however the

geostrophic departures are primarily rotational. The relative vorticity is neglected in the divergence term of the vorticity equation, and the twisting and vertical vorticity advection terms are omitted. These latter assumptions are not necessary, and are made only for computational convenience at this point.

In the (x,y,p,t) system, our basic equations are:

Vorticity Equation

$$\frac{\partial \eta}{\partial t} + \mathbf{V}\cdot\nabla\eta + f\nabla\cdot\mathbf{V} = 0 \tag{1}$$

First Law of Thermodynamics

$$\frac{\partial T}{\partial t} + \mathbf{V}\cdot\nabla T + \sigma\omega = Q = \frac{1}{C_p}\frac{\mathrm{d}q}{\mathrm{d}t} \tag{2}$$

Hydrostatic Equation

$$\frac{\partial \psi}{\partial p} = -\frac{RT}{fp} \tag{3}$$

Continuity Equation

$$\nabla\cdot\mathbf{V} + \frac{\partial \omega}{\partial p} = 0 \tag{4}$$

Horizontal Wind

$$\mathbf{V} \approx \mathbf{K}x\nabla\psi \tag{5}$$

Here $\mathbf{V}$ is the horizontal wind vector; $\omega \equiv \mathrm{d}p/\mathrm{d}t$; T = temperature; $f = 2\Omega \sin\varphi$ = Coriolis parameter; η = absolute vorticity = $\mathbf{K}\cdot\nabla x\mathbf{V} + f$; σ = static stability = $T/\theta\ \partial\theta/\partial p$; θ = potential temperature; ψ = stream function; C_p = specific heat at constant pressure; $\mathrm{d}q$ = heat added per unit mass; Q = rate of temperature change due to the addition of heat; (x,y,p) are the horizontal and vertical coordinates of a rectangular right-handed coordinate system; p = pressure; t = time; R is the gas constant.

This is a system of five equations in the five unknowns $\mathbf{V}$, ψ, T, η, ω. It has been assumed that the stream function ψ satisfies the hydrostatic equation, and also that Q is known.

Since

$$\psi = \frac{\phi}{f} + \psi^1 \tag{6}$$

where $\phi = gz$ is the geopotential; g = acceleration of gravity; z = geometrical height,

$$\frac{\partial \psi}{\partial p} = \frac{1}{f}\frac{\partial \phi}{\partial p} + \frac{\partial \psi^1}{\partial p} = -\frac{RT}{fp} + \frac{\partial \psi^1}{\partial p} \tag{6a}$$

due to the hydrostatic law. It is thus seen that assumption (3) implies that

$$\psi^1(x,y,p,t) = \psi^1(x,y,t), \tag{6b}$$

i.e. the geostrophic deviations are invariant with height.

B. Modelling Approximations

In principle, it is possible to apply the system of equations (1)–(5) inclusive at any number of levels in order to solve the problem. Since the vorticity equation (1) becomes a second order partial differential equation in $\partial\psi/\partial t$, it is desirable from the computational point of view to avoid solving the system of equations at very many levels. It is possible to simplify the computations by making certain modelling approximations, i.e. certain vertical constraints upon the parameters of the problem.

We have already made one modelling assumption, (6b). In addition it is assumed that the vertical variation of the static stability is known, i.e.

$$\sigma(x,y,p,t) = \frac{T}{\theta}\frac{\partial \theta}{\partial p} = \frac{\partial T}{\partial p} - \frac{\kappa T}{p} = F(p)\bar{\sigma}(x,y,t) \tag{7}$$

where the mean value of a quantity X is defined by

$$\bar{X} = \frac{1}{p_0 - \epsilon}\int_{\epsilon}^{p_0} X\,\mathrm{d}p \tag{8}$$

and $p_0 = 1000$ mb, $\epsilon = 0{\cdot}1$ mb, $\kappa = R/C_p = 0{\cdot}288$.

From equation (7), it is seen that

$$\frac{\partial}{\partial p}(Tp^{-\kappa}) = F\bar{\sigma}p^{-\kappa} \tag{9}$$

$$T(x,y,t) = \bar{\sigma}(x,y,t)\,p^{\kappa}\int Fp^{-\kappa}\mathrm{d}p + p^{\kappa}g(x,y,t)$$

where $g(x,y,t)$ is an arbitrary function.

If the bar operator (8) is applied to equation (9) and the result subtracted from (9)

$$T(x,y,p,t) = A(p)\bar{\sigma}(x,y,t) + B(p)\bar{T}(x,y,t) \quad (10)$$

where

$$A(p) = p^{\kappa}\int F(p)p^{-\kappa}\mathrm{d}p - (1+\kappa)\left(\frac{p}{p_0}\right)^{\kappa}\overline{[p^{\kappa}\int F(p)p^{-\kappa}\mathrm{d}p]} \quad (11)$$

$$B(p) = (1+\kappa)\left(\frac{p}{p_0}\right)^{\kappa}$$

We may now apply equation (10) at $p = p_0$ and $p = \epsilon$ and solve the resulting pair of equations for $\bar{\sigma}$ and $\bar{T}$ in terms of T_{p_0} and T_ϵ. Substituting these values back into equation (10), we find

$$T(x,y,p,t) = G(p)T_{p_0}(x,y,t) + H(p)T_\epsilon(x,y,t) \quad (12)$$

where

$$\left.\begin{aligned} G(p) &= \frac{\left(\frac{\epsilon}{p_0}\right)^{\kappa}\frac{A(p)}{A(p_0)} - \left(\frac{p}{p_0}\right)^{\kappa}\frac{A(\epsilon)}{A(p_0)}}{\left(\frac{\epsilon}{p_0}\right)^{\kappa} - \frac{A(\epsilon)}{A(p_0)}} \\ H(p) &= \frac{\left(\frac{p}{p_0}\right)^{\kappa} - \frac{A(p)}{A(p_0)}}{\left(\frac{\epsilon}{p_0}\right)^{\kappa} - \frac{A(\epsilon)}{A(p_0)}} \end{aligned}\right\} \quad (13)$$

and hence

$$G(p) = \left(\frac{p}{p_0}\right)^{\kappa}\left[1 - \left(\frac{\epsilon}{p_0}\right)^{\kappa}H(p)\right]$$

We have derived equation (12) in order to eliminate the necessity for predicting $\bar{\sigma}$ and $\bar{T}$. In order to predict $\bar{T}$ from the First Law of Thermodynamics, a value for ω is required at the mean level. If we predict T_ϵ and T_0, instead of $\bar{T}$ and $\bar{\sigma}$, and assume as boundary conditions

$$\left.\begin{aligned} \omega &= 0 \text{ at } p = p_0 \\ \omega &= 0 \text{ at } p = \epsilon \end{aligned}\right\} \quad (14)$$

then there is no need to solve for ω in order to predict the temperature changes at various levels. This is not to say, however, that interior vertical motions do not contribute to temperature changes. It may be seen from equation (1), after substituting from equation (4), that internal vertical motion and divergence will affect the vorticity at all levels, and this will in turn affect the horizontal wind, which affects the horizontal temperature advection.

From equation (3) we see

$$\psi(p) - \psi(p_0) = \frac{R}{f}\int_p^{p_0}\frac{T}{p}\,\mathrm{d}p \quad (15)$$

$$\bar{\psi} - \psi(p_0) = \overline{\frac{R}{f}\int_p^{p_0}\frac{T}{p}\,\mathrm{d}p} \quad (16)$$

$$\psi = \bar{\psi} + \frac{R}{f}\left(\int_p^{p_0}\frac{T}{p}\,\mathrm{d}p - \overline{\int_p^{p_0}\frac{T}{p}\,\mathrm{d}p}\right) \quad (17)$$

substituting equation (12) into equation (17) we get

$$\psi(x,y,p,t) = \bar{\psi}(x,y,t) + I(p)T_{p_0}(x,y,t) + J(p)T_\epsilon(x,y,t) \quad (18)$$

where

$$\left.\begin{aligned} I(p) &= \frac{R}{f}\left[\int_p^{p_0}\frac{G}{p}\,\mathrm{d}p - \overline{\int_p^{p_0}\frac{G}{p}\,\mathrm{d}p}\right] \\ J(p) &= \frac{R}{f}\left[\int_p^{p_0}\frac{H}{p}\,\mathrm{d}p - \overline{\int_p^{p_0}\frac{H}{p}\,\mathrm{d}p}\right] \\ I(p) &= \frac{R}{f\kappa}\left[\left(\frac{p}{p_0}\right)^{\kappa} - \frac{1}{1+\kappa}\right] - \left(\frac{\epsilon}{p_0}\right)J(p) \end{aligned}\right\} \quad (19)$$

We may now write

$$\mathbf{V} = \mathbf{K}\times\nabla[\bar{\psi} + I(p)T_{p_0} + J(p)T_\epsilon] \quad (20)$$

$$\eta = \nabla^2[\bar{\psi} + I(p)T_{p_0} + J(p)T_\epsilon] + f \quad (21)$$

C. Derivation of Final Equations

For the sake of convenience, we introduce a new variable

$$\tau = \frac{RT}{f} \quad (22)$$

into equations (2), (3), (20) and (21). If we now apply the bar operator to equation (1), apply equation (2) at $p = p_0$ and $p = \epsilon$, and introduce equations (20) and (21), we obtain

$$\left.\begin{aligned} \nabla^2\frac{\partial\bar{\psi}}{\partial t} &= J^*(\nabla^2\bar{\psi} + f, \bar{\psi}) + \overline{I_1^2}J^*(\nabla^2\tau_{p_0}, \tau_{p_0}) + \\ &+ \overline{I_1J_1}J^*(\nabla^2\tau_\epsilon, \tau_{p_0}) + \overline{I_1J_1}J^*(\nabla^2\tau_{p_0}, \tau_\epsilon) + \\ &+ \overline{J_1^2}J^*(\nabla^2\tau_\epsilon, \tau_\epsilon) \end{aligned}\right\} \quad (23)$$

$$\frac{\partial \tau_{p_0}}{\partial t} = J^*[\tau_{p_0}, \bar{\psi} + I_1(p_0)\tau_\epsilon] - \frac{\tau_{p_0}}{f} J^*[\bar{\psi} + I_1(p_0)\tau_{p_0} + J_1(p_0)\tau_\epsilon, f] \quad (24)$$

$$\frac{\partial \tau_\epsilon}{\partial t} = J^*[\tau_\epsilon, \bar{\psi} + I_1(\epsilon)\tau_{p_0}] - \frac{\tau_\epsilon}{f} J^*[\bar{\psi} + I_1(\epsilon)\tau_{p_0} + I_1(\epsilon)\tau_\epsilon, f] + Q_{1_\epsilon} \quad (25)$$

where

$$\left.\begin{aligned} I_1 &= \frac{f}{R} I \\ J_1 &= \frac{f}{R} J \\ Q_{1_\epsilon} &= \frac{R}{f} Q_\epsilon \end{aligned}\right\} \quad (26)$$

$J^*(a,b)$ = Jacobian of (a,b) with respect to (x, y)

We observe that the system (23)–(25) is complete. Given initial values of $\bar{\psi}$, τ_{p_0}, τ_ϵ, it is possible to solve for $\partial\bar{\psi}/\partial t$, $\partial\tau_{p_0}/\partial t$, $\partial\tau_\epsilon/\partial t$ and iterate in time to obtain new values. It is assumed that Q_{1_ϵ} is known. The coefficients $\overline{I_1^2}$, $\overline{J_1^2}$, $\overline{I_1 J_1}$, $I_1(p_0)$, $J_1(p_0)$, $I_1(\epsilon)$, $J_1(\epsilon)$ are to be determined empirically. A detailed discussion of these coefficients is given in a later section.

The vorticity equation (23) states that the vorticity of the mean motion may be changed not only by advection of mean absolute vorticity by the mean wind (i.e. the barotropic term), but also by baroclinic "development" terms. These are, advection of thermal vorticity at the surfaces $p = p_0$, ϵ by the thermal winds at $p = p_0$, ϵ; advection of thermal vorticity at $p = p_0$, ϵ by the thermal winds at $p = \epsilon$, p_0 respectively. Thus the extent to which upper level heating will influence the mean motion will depend mainly upon the magnitude of the horizontal temperature gradients which are established by the heating.

The temperature equations (24) and (25) simply state that the temperature at $p = p_0$ changes only by horizontal advection, while the temperature at $p = \epsilon$ is governed by horizontal advection and heating. The second Jacobian on the right side of each equation arises because we have allowed $1/f$ to vary in the definition (22) of τ. We note that, even though no heating takes place directly at $p = p_0$, the temperature at $p = \epsilon$ enters in this equation, so that the indirect effects of upper heating are felt in the surface temperature equation. A similar indirect effect occurs in the vorticity equation.

The fact that we have used the boundary conditions (14), together with the type of formulation made here, has enabled us to derive a prediction system which requires no explicit knowledge of the vertical motion. However, it is desirable to be able to compute this quantity. Usually, a second order partial differential equation for the vertical motion is derived by combining the vorticity equation and the First Law of Thermodynamics. For this model, however, an algebraic equation for ω exists. The use of such an equation, which will now be derived, obviously permits a large simplification in the solution of the problem.

We return to the vorticity equation (1) with (4) used in the divergence term:

$$\frac{\partial \eta}{\partial t} + \mathbf{V}\cdot\nabla\eta - f\frac{\partial \omega}{\partial p} = 0 \quad (27)$$

Applying the bar operator

$$\frac{\partial \bar{\eta}}{\partial t} + \overline{\mathbf{V}\cdot\nabla\eta} = 0 \quad (28)$$

From equations (21) and (22)

$$\frac{\partial \eta}{\partial t} = \frac{\partial \bar{\eta}}{\partial t} + I_1(p)\nabla^2\frac{\partial \tau_{p_0}}{\partial t} + J_1(p)\nabla^2\frac{\partial \tau_\epsilon}{\partial t} \quad (29)$$

Therefore

$$(\mathbf{V}\cdot\nabla\eta - \overline{\mathbf{V}\cdot\nabla\eta}) + I_1(p)\nabla^2\frac{\partial \tau_{p_0}}{\partial t} + J_1(p)\nabla^2\frac{\partial \tau_\epsilon}{\partial t} = f\frac{\partial \omega}{\partial p} \quad (30)$$

i.e.

$$\omega = -\frac{1}{f}\left\{\int_p^{p_0}(\mathbf{V}\cdot\nabla\eta - \overline{\mathbf{V}\cdot\nabla\eta})\,dp + \nabla^2\left[\int_p^{p_0} I_1\,dp\,\frac{\partial \tau_{p_0}}{\partial t} + \int_p^{p_0} J_1\,dp\,\frac{\partial \tau_\epsilon}{\partial t}\right]\right\} \quad (31)$$

Substituting for $(\mathbf{V}\cdot\nabla\eta - \overline{\mathbf{V}\cdot\nabla\eta})$,

$$\omega = -\frac{1}{f}\left\{\begin{aligned}&J^*[\bar{\psi}, \nabla^2\left(\textstyle\int_p^{p_0} I_1 \mathrm{d}p\tau_{p0} + \int_p^{p_0} J_1 \mathrm{d}p\tau_\epsilon\right)] + \\ &+ J^*\left[\textstyle\int_p^{p_0} I_1 \mathrm{d}p\tau_{p0} + \int_p^{p_0} J_1 \mathrm{d}p\tau_\epsilon, \nabla^2\bar{\psi} + f\right] + \\ &+ \left[\textstyle\int_p^{p_0} I_1^2 \mathrm{d}p - (p_0 - p)\overline{I_1^2}\right] J^*(\tau_{p0}, \nabla^2\tau_{p0}) + \\ &+ \left[\textstyle\int_p^{p_0} I_1 J_1 \mathrm{d}p - (p_0 - p)\overline{I_1 J_1}\right] J^*(\tau_{p0}, \nabla^2\tau_\epsilon) + \\ &+ \left[\textstyle\int_p^{p_0} I_1 J_1 \mathrm{d}p - (p_0 - p)\overline{I_1 J_1}\right] J^*(\tau_\epsilon, \nabla^2\tau_{p0}) + \\ &+ \left[\textstyle\int_p^{p_0} J_1^2 \mathrm{d}p - (p_0 - p)\overline{J_1^2}\right] J^*(\tau_\epsilon, \nabla^2\tau_\epsilon) + \\ &+ \nabla^2\left[\int_p^{p_0} I_1 \mathrm{d}p \frac{\partial\tau_{p0}}{\partial t} + \int_p^{p_0} J_1 \mathrm{d}p \frac{\partial\tau_\epsilon}{\partial t}\right]\end{aligned}\right\} \tag{32}$$

We observe the following features of this algebraic vertical motion equation: the variation of ω with height depends upon I_1, J_1, and certain combinations of them. These functions are related to $G(p)$ and $H(p)$, which are related to $A(p)$ and ultimately to $F(p)$. Thus the variation of ω with height depends explicitly on the variation of static stability with height. One is therefore not at liberty to select a parabolic variation of ω, as is very frequently done. We also notice that, as long as there exists a temperature gradient either at $p = \epsilon$ or $p = p_0$, $\omega \neq 0$ necessarily. But if both temperature gradients are zero, ω will be identically zero, unless the variation of $1/f$ is included. What this says is that, if the temperature remains constant everywhere, the fluid is thermally inactive, and behaves like a barotropic fluid.

The experiment currently under way treats the heating case in the following manner: it is assumed that the temperature at $p = \epsilon = 0{\cdot}1$ mb is constant initially. It is then assumed that asymmetrical heating takes place statically, and that the motion itself is not affected for a finite period of time while the temperature is changed. That is,

$$\frac{\partial T_\epsilon}{\partial t} = Q_\epsilon, \quad t_1 \leqslant t \leqslant 0 \tag{33}$$

Such heating is allowed to take place for 24 hr and then the heat source is removed. Subsequent temperature changes are governed by equation (25) with $Q_{1\epsilon} = 0$. New values of $I_1(\epsilon)$, $J_1(\epsilon)$ in terms of the new temperature distribution are necessary and all of the other coefficients will have new values as well.

A complete statement of the prognostic problem is

$$\nabla^2\frac{\partial\bar{\psi}}{\partial t} = J^*(\nabla^2\bar{\psi} + f, \bar{\psi}) + \overline{I_1^2} J^*(\nabla^2\tau_{p0}, \tau_{p0}) + \\ + \overline{I_1 J_1} J^*(\nabla^2\tau_\epsilon, \tau_{p0}) + \overline{I_1 J_1} J^*(\nabla^2\tau_{p0}, \tau_\epsilon) + \\ + \overline{J_1^2} J^*(\nabla^2\tau_\epsilon, \tau_\epsilon) \tag{23}$$

$$\frac{\partial\tau_{p0}}{\partial t} = J^*[\tau_{p0}, \bar{\psi} + I_1(p_0)\tau_\epsilon] - \\ \frac{\tau_{p0}}{f} J^*[\bar{\psi} + I_1(p_0)\tau_{p0} + J_1(p_0)\tau_\epsilon, f] \tag{24}$$

$$\frac{\partial\tau_\epsilon}{\partial t} = Q_{1\epsilon}, \quad -24 \leqslant t \leqslant 0 \tag{33}$$

$$\frac{\partial\tau_\epsilon}{\partial t} = J^*[\tau_\epsilon, \bar{\psi} + I_1(\epsilon)\tau_{p0}] - \\ - \frac{\tau_\epsilon}{f} J^*[\bar{\psi} + I_1(\epsilon)\tau_{p0} + J_1(\epsilon)\tau_\epsilon, f], \; t > 0 \tag{25a}$$

III. THE HEATING FUNCTION

The heating function, $q(x,y,p,t)$ is constructed in such a manner that the temperature change due to heating (Q) varies linearly from 0 at $p = 0$ to a maximum at $p = \epsilon$ and then decreases linearly to 0 at $p = 2\epsilon$. That is,

$$Q(x,y,p,t) = \left\{\begin{aligned}&\frac{p}{\epsilon} Q_\epsilon(x,y,t) && , 0 \leqslant p \leqslant \epsilon \\ &\left(-\frac{p}{\epsilon} + 2\right) Q_\epsilon(x,y,t), && \epsilon \leqslant p \leqslant 2\epsilon \\ &0 && , 2\epsilon \leqslant p \leqslant p_0\end{aligned}\right\} \tag{34}$$

where Q_ϵ is the rate of temperature change due to heating at $p = \epsilon$, see equation (3). It can be seen from equation (34) that Q_ϵ is twice the average value of Q in the layer $0 \leqslant p \leqslant 2\epsilon$.

Since the density varies non-linearly with p, the shape of the heating function (q) is similar to that given in Fig. 1.

In what follows, we shall specify the heating function over the entire region of integration, which is taken to be that portion of the northern hemisphere contained within the 30° N latitude

circle. The 30° latitude circle is chosen as the southern boundary of the region for the following reasons: inasmuch as we are attempting to simulate the effects of auroral heating, we wish to confine ourselves to those regions in which auroral activity is greatest (i.e. in the vicinity of 60° N). However, occasional aurorae extend much farther south.

phenomena. It corresponds to about 1/10,000 of the normal solar heating over the region and was obtained by assuming a rate of heat input of 10^{-2} cal cm^{-2} min^{-1} over a ring 100 km wide centered at 60° N. The actual heat input at each point is assumed to be proportional to the observed auroral frequency.[4] With this average heating rate, $Q_\epsilon(x,y)$ the rate of temperature

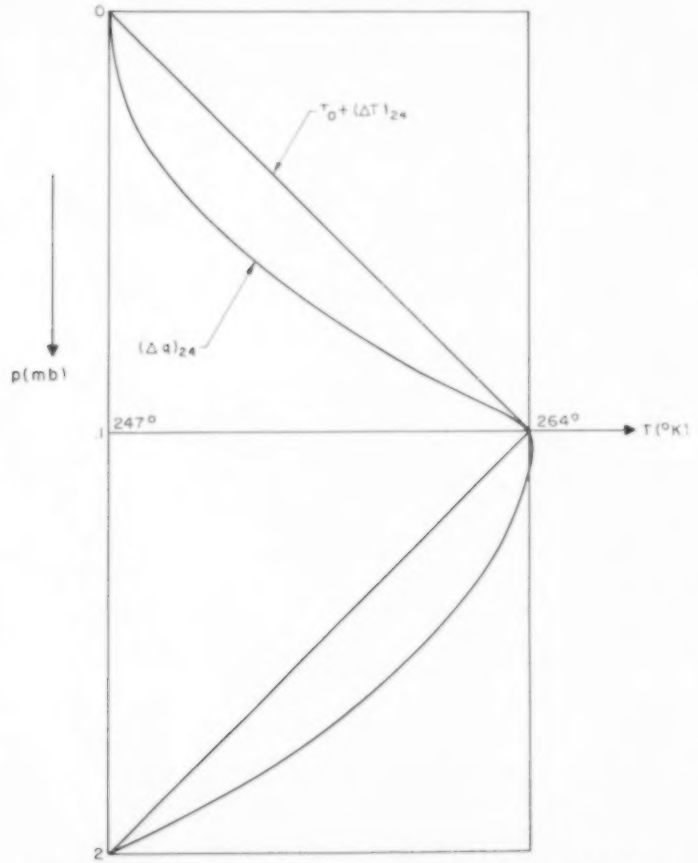


Fig. 1. Schematic diagram of variation of heating function with height.

We need to determine Q_ϵ in order to solve equation (33). To do this, we assume that the average heat input over the entire region is known, and is equal to $1{\cdot}570 \times 10^{-4}$ cal cm^{-2} min^{-1}.

This rate of heating is a conservative estimate based upon an intensity of solar corpuscular emission consistent with observed auroral

change due to heating at each point at the level $p = \epsilon$ is given by,

$$Q_\epsilon(x,y) = 0{\cdot}535 \times 10^{-5} I(x,y) \text{ deg sec}^{-1} \quad (35)$$

where $I(x,y)$ is the distribution of auroral frequencies. The function $Q_\epsilon(x,y)$ is shown in Fig. 2.

Fig. 1 is a schematic diagram showing typical

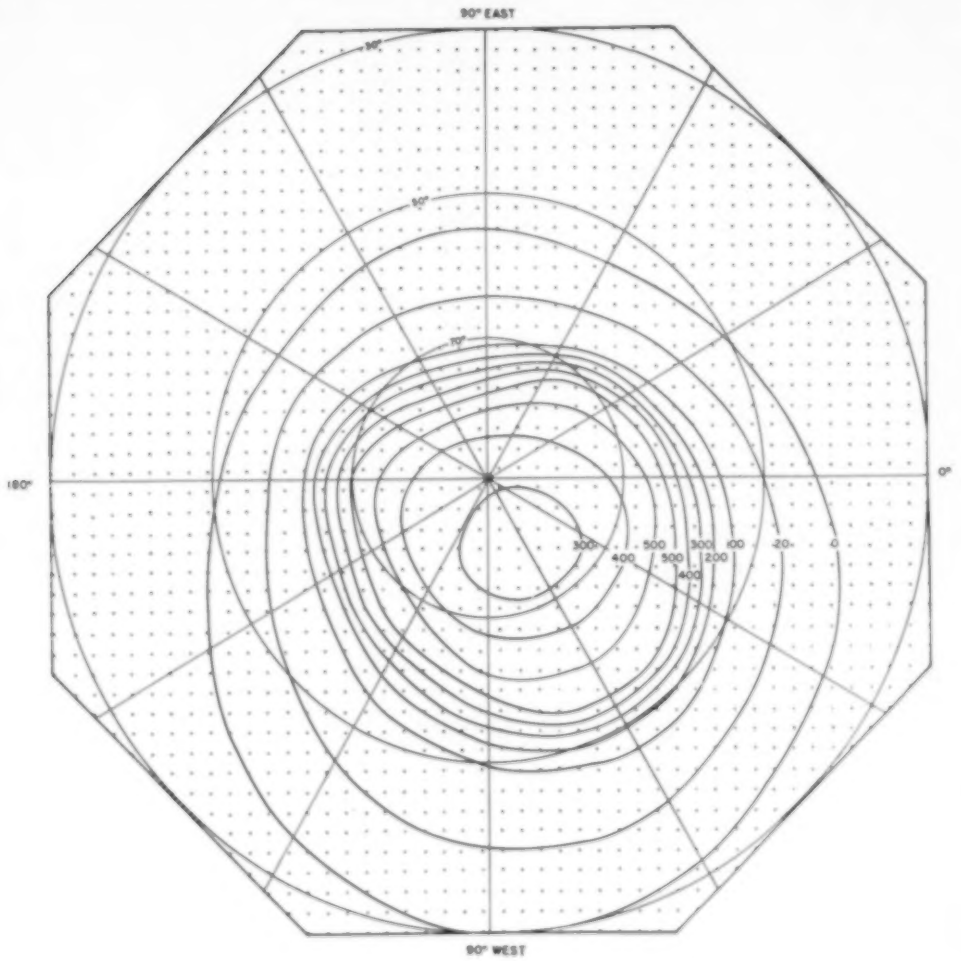


Fig. 2. Heating function and grid points.

temperature changes in the layer $0 \leq p \leq 2\epsilon$ at a point corresponding to a maximum of auroral frequency.

IV. THE INITIAL CONDITIONS

In the first experiment, we shall assume analytical initial conditions. Since we have three independent prognostic equations, we have to specify three fields, $\bar{\psi}$, τ_{p_0}, τ_s, independently.

To determine the $\bar{\psi}$, τ_{p_0}, τ_s fields, we shall in each case assume that no perturbation exists initially. Thus the stream field will be such that only an initial zonal flow exists. The aim of the problem will be to see whether any perturbation develops on this zonal flow.

We assume that, on the spherical Earth, the basic current of the vertically integrated flow U, is given by

$$\bar{U} = a\alpha \sin 2q = -\frac{\partial \bar{\psi}}{\partial y} = -\frac{\partial \bar{\psi}}{a \partial q} \tag{36}$$

where a = radius of the Earth
α = angular velocity of the air
q = latitude.

If we write as a boundary condition

$$\bar{\psi} = \bar{\psi}_0 \text{ at } q = q_0 = 30°$$

then integration of equation (36) gives

$$\bar{\psi} = \bar{\psi}_0 + \frac{a^2 a}{2}(\cos 2q - 0{\cdot}500) \qquad (37)$$

If we assume $\bar{\psi}_0 \sim g z_0/f$ cm^2 sec^{-1},

z_0 = height at $q = 30^\circ$ N = 6 km,

and $a = 3 \times 10^{-6}$ sec^{-1}, corresponding to $U \approx 15$ m sec^{-1} at 45° N latitude, then

$$\bar{\psi} = (7{\cdot}909 + 0{\cdot}304 \cos 2q) \times 10^{12} \text{ cm}^2 \text{ sec}^{-1} \qquad (38)$$

To determine the τ_{p_0} field, we assume that τ_{p_0} is given by

$$\frac{\partial T_{p_0}}{\partial y} = \frac{\partial T_{p_0}}{a \partial q} = -k \sin 2q \qquad (39)$$

where $-k$ = maximum amplitude of the temperature gradient at $p = p_0$. With the boundary condition $(T_{p_0}) = T_{30}$ at $q = 30^\circ$ N,

$$T_{p_0} = \frac{ak}{2}(\cos 2q - 0{\cdot}500) + T_{30} \qquad (40)$$

Furthermore, if $(T_{p_0})_{90} = T_{90}$ at $q = 90^\circ$ N,

$$T_{p_0} = \frac{ak}{2}(1 + \cos 2q) + T_{90} \qquad (41)$$

Since the two T_{p_0}'s must be equal,

$$k = \frac{T_{30} - T_{90}}{\frac{3a}{4}} \qquad (42)$$

If we choose $T_{30} = 288^\circ A$, $T_{90} = 258^\circ A$, then

$$\left.\begin{aligned} k &= 0{\cdot}626^\circ\text{C}/100 \text{ km} \\ T_{p_0} &= 19{\cdot}931 \cos 2q + 278{\cdot}034 \end{aligned}\right\} \qquad (43)$$

Then

$$\tau_{p_0} = \frac{7{\cdot}979 + 0{\cdot}572 \cos 2q}{1{\cdot}458 \sin q} \times 10^{12} \text{ cm}^2 \text{ sec}^{-1} \qquad (44)$$

We shall assume that the initial temperature field at $p = \epsilon$ in the no heating case is given by a constant, which according to Murgatroyd[5], if $\epsilon = 0{\cdot}1$ mb, gives

$$T_\epsilon = 247^\circ A \qquad (45)$$

In terms of τ_ϵ (which varies due to $1/f$),

$$\tau_\epsilon = \frac{7{\cdot}089}{1{\cdot}458 \sin q} \times 10^{12} \text{ cm}^2 \text{ sec}^{-1} \qquad (46)$$

After heating takes place τ_ϵ is governed by a different relationship according to equation (33) and will be discussed in greater detail in a later section.

V. THE COEFFICIENTS OF THE EQUATIONS

It is necessary to determine the coefficients of the prognostic system empirically.

Let us define

$$\bar{\bar{T}} = \frac{1}{At}\iiint T \, dx \, dy \, dt \qquad (47)$$

where A is area, t is time, so that $\bar{\bar{T}}$ is a time average over an area. From normal temperature maps it is possible to determine a $\bar{\bar{T}}$ for several pressure levels. If the operator (47) is applied to definition (10), we have

$$A(p) = \frac{\bar{\bar{T}}(p) - B(p)\,\tilde{\bar{T}}}{\tilde{\bar{\sigma}}} \qquad (48)$$

To get $\bar{\bar{T}}(p)$, the following sources were used: The mean sea level temperature for January from the Handbook of Meteorology[6]; temperatures from 685–111 mb inclusive from the Handbook of Geophysics[7]; temperatures from 100–0·1 mb inclusive, from Murgatroyd[5]. These profiles are shown in Figs. 3 and 4.

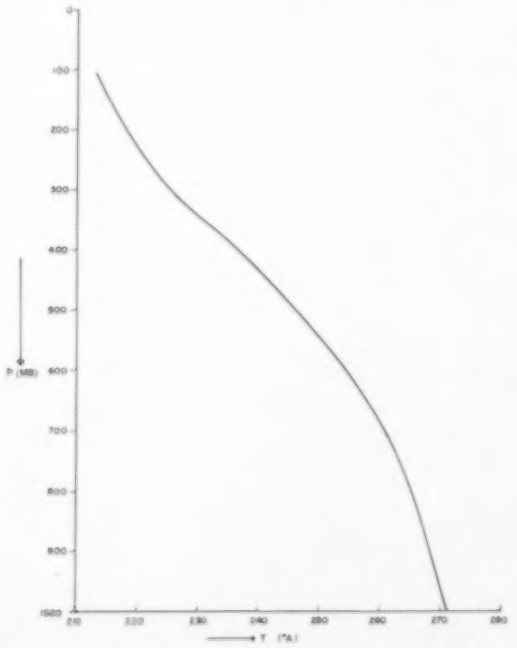


Fig. 3. Variation of average January temperature with height.

$\tilde{\sigma}$ may be easily computed from the finite difference form of

$$\tilde{\sigma} = \frac{\partial \hat{T}}{\partial p} - \frac{\kappa \hat{T}}{p} \tag{49}$$

Once $A(p)$ is computed, G and H are easily found from equation (13), and I and J from equation (19). The profiles G,H,I,J are shown in Figs. 5 and 6. Here we display these profiles only for the no-heating case. They will change slightly if the temperature profile after heating, as shown in Fig. 4, is used.

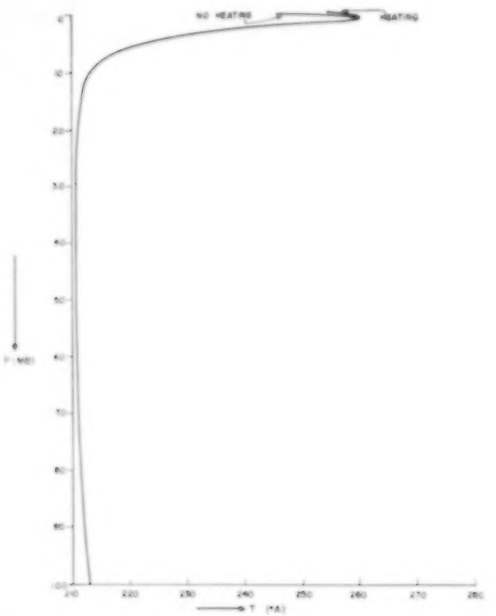


Fig. 4. Variation of average January temperature with height.

From the G and H profiles and equation (12) it is seen that, in this model, the influence of T_{p_0} and T_ϵ are equal at a relatively low value of pressure (40 mb). Below this level the influence of T_{p_0} is dominant, and above this level the influence of T_ϵ is dominant and rapidly increasing.

According to equation (20), the function J, which describes the effect of the thermal wind at $p = \epsilon$ on winds at lower levels is rather small up to about 150 mb, and only becomes quite dominatingly large near the very top of the

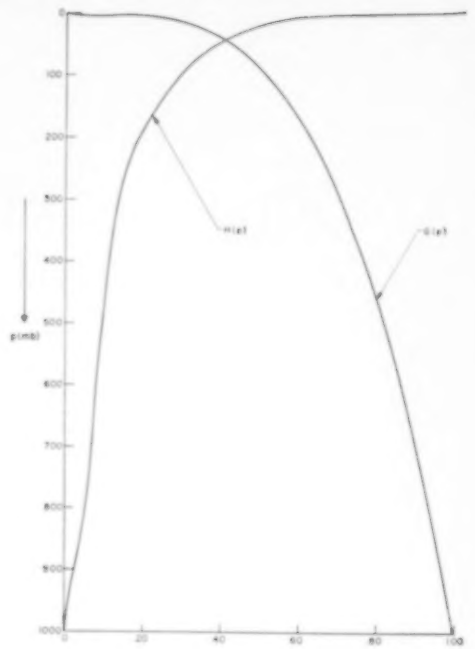


Fig. 5. Functions describing variation of model temperature with height.

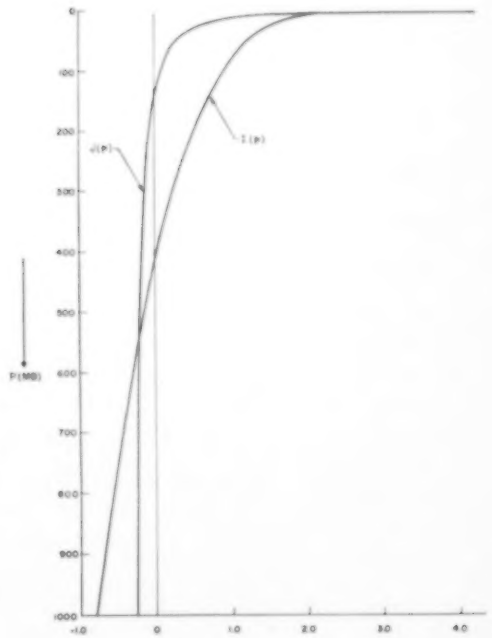


Fig. 6. Functions describing variation of model wind and vorticity with height.

atmosphere. I, on the other hand, has a reasonably large effect at all levels except from about 350–550 mb, where both functions are small. This implies that, in this region, the wind is approximately given by the mean wind.

VI. METHOD OF SOLUTION

The finite-difference form of the system (23), (24), (25a) is being programmed for solution on the IBM 704 computer. The 30° latitude circle is approximated by a regular octagon externally tangent to the 30° latitude circle on a polar stereographic projection. The region is then divided up into a mesh containing 1752 equally spaced grid points (see Fig. 2).

The computations will proceed in the following manner:

(a) Evaluate $\bar{\psi}$, τ_{p_0}, τ_ϵ at the grid points from equations (38), (44), (46).

(b) Solve the finite-difference forms of equations (23), (24), (25a) for

$$\frac{\partial \bar{\psi}}{\partial t}, \quad \frac{\partial \tau_{p_0}}{\partial t}, \quad \frac{\partial \tau_\epsilon}{\partial t}.$$

(c) Step these forward in time by the following iteration scheme

$$(\bar{\psi}, \tau_{p_0}, \tau_\epsilon)_t = (\bar{\psi}, \tau_{p_0}, \tau_\epsilon)_{t-2} + 2\Delta t \left(\frac{\partial \bar{\psi}, \tau_{p_0}, \tau_\epsilon}{\partial t}\right)_{t-1}$$

after the first time step. Forward difference formulas are used for the first time step.

(d) Repeat the process by returning to step (a), for as many time steps as desired.

In order to avoid computational instability it is necessary in this problem to use $\Delta t = 40$ minutes.[8] This allows for maximum winds up to about 70 m sec^{-1}. The vertical motion ω will be computed at $p = 900, 700, 500, 300, 100$ mb every few time steps, from the finite-difference analogue of equation (32).

In order to have some further control over the stability of the computations, it is planned to compute the total (potential plus kinetic plus internal) energy change per unit time over the volume under consideration. The differential system conserves energy, and it is desirable to preserve this constraint in the finite-difference system. The boundary conditions are such as to permit no inflow or outflow of energy through the walls of the system.

The preceding discussion applies without modification to the predictions in the heating case. The only difference will be in τ_ϵ at $t = 0$. In the heating case, the τ_ϵ distribution will be that corresponding to the original constant temperature field which is altered by statically heating for 24 hr by means of the asymmetrical heating function described in Section III.

Calculations will be made both with and without the heating, and the development of perturbations in the flow field in mid-troposphere in response to the high level heating will be studied. The differences in the amplitudes of the perturbations between the heating and no-heating cases will be a measure of the effect of the heating. It is realized that truncation error may grow after many time steps have elapsed. However, truncation error should be almost identical for both the heating and no-heating cases. Consequently, if the computations remain stable in spite of accumulated truncation error, differences in truncation error may be neglected, so that we will indeed be studying a true effect. We shall also be able to compute the rate of change of potential and kinetic energy at each point, so that it will be possible to compute the rate of energy propagation downward.

VII. FUTURE PLANS

In the present experiment, we have taken as an initial heating function that which might exist a short time period after the outbreak of corpuscular radiation, i.e. we have allowed the energy which entered the atmosphere within a narrow ring to be diffused in such a manner that the total energy remained constant and was distributed in proportion to area.

In our next experiment, we plan to start the integrations at the moment the intense concentration of radiation reaches auroral level, so that the energy per unit area within a narrow ring will be very high. We then plan to follow this energy down by means of transport equations which permit fine resolution in the vertical.

REFERENCES

1. W. NUPEN and M. KAGEORGE, Bibliography on Solar-Weather Relationships in *Meteorological Abstracts and Bibliography*, American Meteorological Society, Contract AF19(604)-2062 (1958).
2. R. SHAPIRO, *J. Met.* **13**, 335 (1956).
3. R. SHAPIRO, *J. Met.* —, — (1959).
4. L. HARANG, pp. 347–355 in *Compendium of Meteorology* (ed. by T. F. MALONE), American Meteorological Society, Boston, Mass. (1951).
5. R. J. MURGATROYD, *Quart. J. R. Met. Soc.* **83**, 417 (1957).
6. F. A. BERRY, JR., E. BOLLAY and N. R. BEERS, p. 949 in *Handbook of Meteorology*, McGraw-Hill, New York (1945).
7. *Handbook of Geophysics*, pp. 223–240, Geophysics Research Directorate, Air Force Cambridge Research Center (1957).
8. J. G. CHARNEY, R. FJORTOFT and J. VON NEUMANN, *Tellus* **2**, 237 (1950).

RECENT INFORMATION GAINED FROM SATELLITE ORIENTATION MEASUREMENT

ROBERT J. NAUMANN*
Army Ballistic Missile Agency,
Research Projects Laboratory, Development Operations Division

Abstract—Much attention has been given to the motion of the center of mass of a satellite, but comparatively little attention has been given to the motion of the body about its center of mass. Knowledge of body motion about the mass center and the actual orientation at particular times is necessary to evaluate certain information gained from satellites.

This paper points out certain anomalies in temperature measurements from the Army Explorer IV which suggested that external torques may be acting on the vehicle. This led to radio frequency measurements which confirmed the fact that the angular momentum vector is precessing. This paper also outlines a method whereby the orientation of a vehicle may be determined at any desired time.

INTRODUCTION

The problem of determining the spatial orientation of a satellite becomes quite important in evaluating the data received. Since the results of most experiments that have been and will be flown depend on the direction in which a particular detector is looking at a particular time, a correct evaluation of such data can be obtained only if the spatial orientation of the detector is known. This is particularly true for radiation experiments or any other experiments where the object to be measured has a preferred direction in space. For the case of an asymmetrical payload, the orientation with respect to the Sun will strongly affect the temperature of the interior. Also, the orientation with respect to the velocity vector will affect the lifetime and atmospheric density measurements. If we assume that the satellite was spun before injection into orbit, such as in the case of the Explorers, one would certainly expect, through conservation of angular momentum, that the spin vector would remain in a space-fixed direction. Any deviation from this direction indicates the presence of some disturbing torques. Thus, by analysing the change of orientation of a satellite, we can get some idea of the torques acting upon it.

I have chosen for my subject, Explorer IV, the reasons being: (1) Many field strength recordings from various tracking stations are available; (2) a complete ephemeris has been prepared by the Smithsonian Astrophysical Observatory; (3) much radiation data from the Van Allen belt and the Argus effect were obtained from this satellite; (4) an interesting temperature anomaly which could be due to orientation changes was noted.

TRANSITION TO FLAT SPIN

Since we are dealing with Explorer IV, it would be well to mention the transition to flat spin which is common to all the early Explorer satellites. As may be seen in Fig. 1, the satellite was initially spun about its long axis which is an axis of least moment of inertia. If any perturbing forces act, the satellite will begin to precess about the angular momentum vector. The precessing frequency can be found by means of a Poinsot construction, and is given by the total angular momentum divided by the maximum moment of inertia. If angular momentum is a constant of the motion, then the precessing frequency is also a constant of the motion and is given by the ratio of the minimum to maximum moments of inertia times the initial spin. For Explorer IV, the moment of inertia ratio was 1/84; the initial spin was 12 c/s. Therefore, the precessing period should be 7 sec. This is the same as the observed period, which was deduced

* Now with Research Projects Division, Marshall Space Flight Center, NASA, Huntsville, Ala.

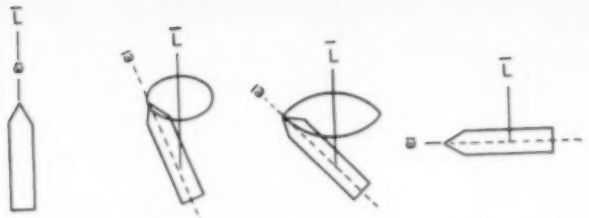


Fig. 1. Evolution of explorer motion. $\bar{L}$ represents the total angular momentum and $\bar{\omega}$ the spin rate about the body symmetry axis.

from field strength recordings and photographic evidence (Fig. 2). Since the exposure for this photograph was 20·6 sec, and since five distinct bright flashes can be seen, we conclude that the satellite has made 2½ cycles in 20 sec, or about 1 cycle each 7 sec. This shows that, even after 3 months, the magnitude of the angular momentum has been conserved.

When the satellite picks up any precession, the flexible portions of the payload are driven in forced oscillation. This causes a loss of energy. Since no dissipative torques have acted on the system, which follows from the fact that the tumble frequency remains constant, $I\omega$ must be conserved while $I\omega^2$ is not. If energy is to be lost at the expense of ω, then I must increase,

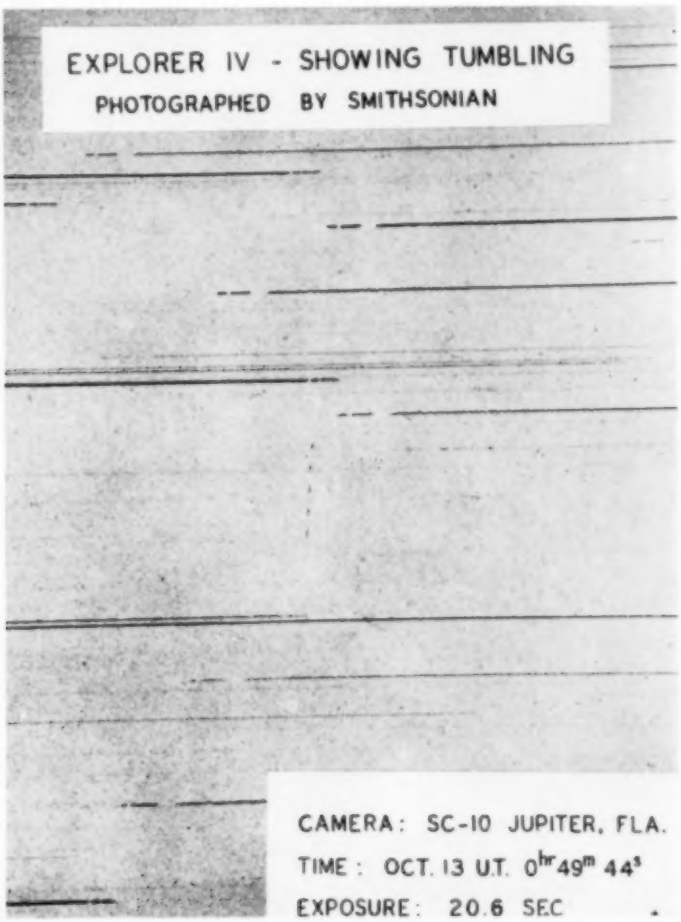

EXPLORER IV - SHOWING TUMBLING
PHOTOGRAPHED BY SMITHSONIAN
CAMERA: SC-10 JUPITER, FLA.
TIME: OCT. 13 U.T. 0hr49m 44s
EXPOSURE: 20.6 SEC

Fig. 2. Explorer IV, showing tumbling.

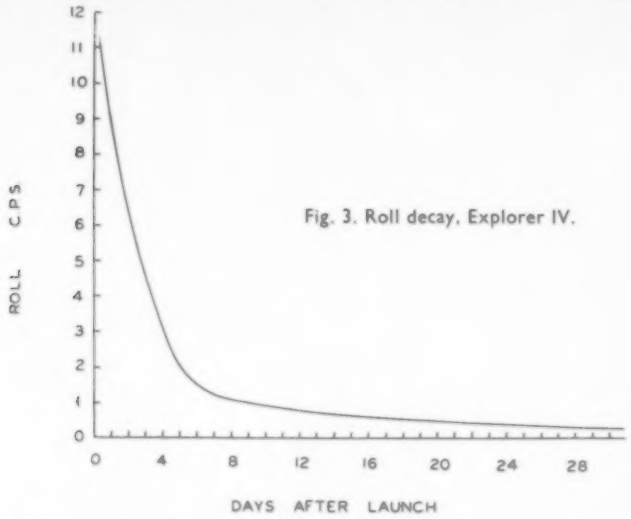


Fig. 3. Roll decay, Explorer IV.

and the body must seek a rotation about an axis of greater moment of inertia. This causes an opening of the cone angle which increases the amplitude of the forced oscillations and results in an even greater opening angle. When the opening angle reaches 90°, the roll has gone to zero and the driving force vanishes. Then the satellite is rotating in a plane normal to its original angular momentum vector. Fig. 3 shows the roll decay, measured from the roll modulation of radiation count rate, from which the opening angle versus time may be calculated. Fig. 4 shows that the angle approaches 90° quite rapidly after a few days.

EVIDENCE FOR CHANGE IN ORIENTATION

If there are no external torques, the plane of tumble should always have the same orientation in space. However, a rather interesting temperature effect which gives evidence to the contrary was noted by Snoddy.[1, 2] Since the thermal properties of the satellite and the orientation of the orbit with respect to the Sun (which determines percentage of time spent in sunlight) are fairly well known, it is possible to construct theoretical temperature curves for a body with various percentages of its area presented to the Sun. Fig. 5 shows such curves with a plot of Explorer IV temperatures superimposed. This

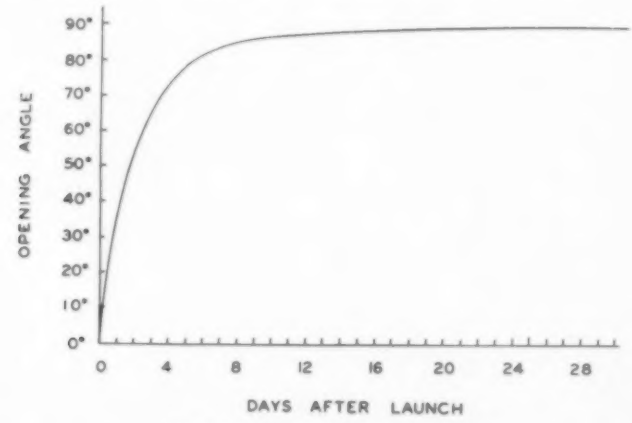


Fig. 4. Opening angle, Explorer IV.

presents strong evidence that the projected area is changing. Since the time-averaged projected area varies from 100 per cent, when the Sun is perpendicular to the plane of tumble, to 63·6 per cent, when the Sun is in the plane of tumble, the variation in temperature indicates an orientation change of about 90° in 10 days. The motion of the orbit about the Sun can account for a maximum change of 1°/day, but certainly not as much as the 9° or 10°/day demanded by temperature measurements.

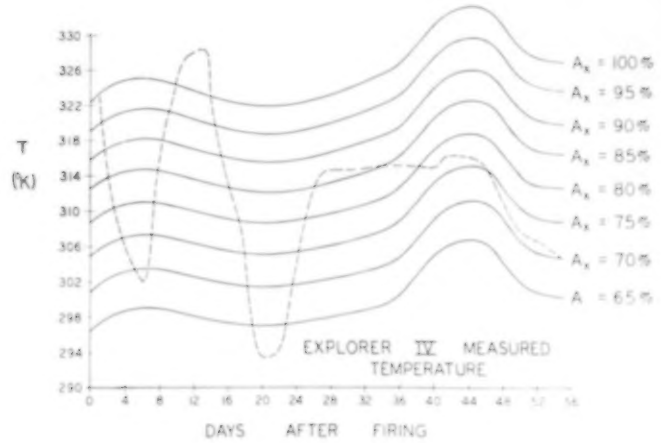


Fig. 5. Measured temperature of Explorer IV superimposed on theoretical curves.

ORIENTATION MEASUREMENT

In an effort to determine whether or not an orientation change was actually responsible for the temperature fluctuations, several methods were devised for using Army Ballistic Missile Agency and Stanford Research Institute radio frequency field strength measurements to determine space-fixed orientation.[3] Since this study was not considered originally, data was taken in a manner that did not readily facilitate such measurements. However, with several successive passes, it was possible to make a "ball park" measurement of orientation by assuming that there was no precession during the passes used for the measurement. Fig. 6 shows the direction of the angular momentum vector projected through the celestial sphere.[4] It may be seen that the precession is not constant in magnitude or direction. As stated, the measurements are crude, being accurate to around $\pm$ 10°, but the presence of external torques is still evident to a degree that cannot be explained by measuring error.

In an effort to define more clearly the nature of these external torques, a more exact method of orientation measurement must be developed. We cannot readily dispense with the assumption that the angular momentum vector remains space-fixed during the time required to measure its direction, but we can reduce the time requirement by measuring the look angle at two different times during one pass. Thus, the time during which we must assume space-fixed orientation is reduced from several hours to several minutes. The look angle ψ is defined as the angle which the line of sight $\mathbf{R}$ makes with the direction of the angular momentum $\boldsymbol{\Omega}$. We can therefore write:

$$\begin{aligned} \mathbf{R}_1 \cdot \boldsymbol{\Omega} &= \cos \psi_1 \\ \mathbf{R}_2 \cdot \boldsymbol{\Omega} &= \cos \psi_2 \\ |\, \boldsymbol{\Omega} \,| &= 1 \end{aligned} \qquad (1)$$

where $\mathbf{R}$ and $\boldsymbol{\Omega}$ are unit vectors.

The calculation of $\mathbf{R}_{1,2}$ is straightforward if the satellite and observer positions are known.

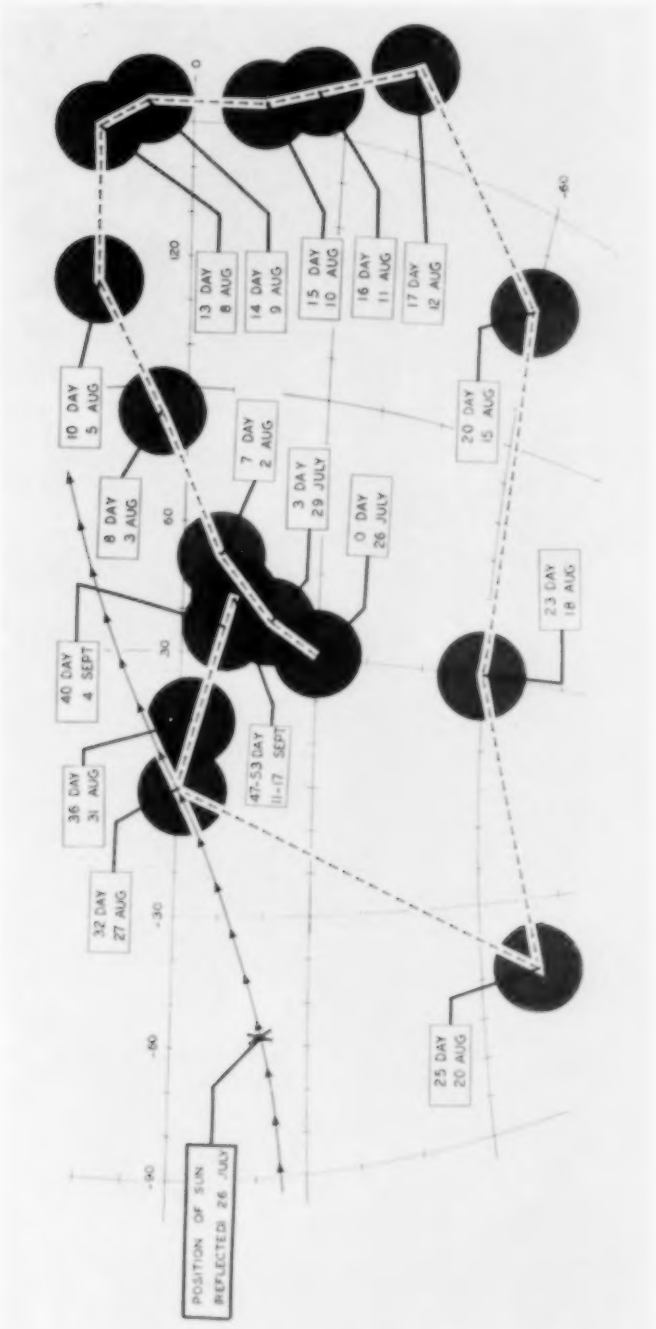


Fig. 6. Projection of Explorer IV angular momentum vector through the celestial sphere.

The problem then becomes that of measuring ψ and using equation (1) to find $\mathbf{\Omega}$.

It should be pointed out that a solar aspect indicator, which will measure the orientation with respect to the Sun, would provide a useful check on field strength measurements. However, the solar aspect angle alone will not provide sufficient data to determine the space-fixed orientation.

To measure ψ, the following technique may be employed. The satellite may be considered an electric dipole, rotating in a plane normal to $\mathbf{\Omega}$. Since a dipole radiates an E wave polarized in the direction of the longitudinal axis, whose magnitude varies as the cosine of the angle the longitudinal axis makes with the line of sight $\mathbf{R}$, the $\mathbf{E}$ will be "seen" by the observer to describe an ellipse in space as the satellite tumbles. The semimajor axis will be $|E|$ and the semiminor axis will be $|E\cos\psi|$.

If the receiver uses a circularly polarized antenna, the ratio of the minimum to the maximum field strength may be measured. This ratio is:

$$\left|\frac{E_{\min}}{E_{\max}}\right| = \left|\frac{E\cos\psi}{E}\right| = \cos\psi \qquad (2)$$

In this manner, the look angle may be measured directly. However, by using separate horizontally and vertically polarized antennas, more information may be obtained. A typical example is shown in Fig. 7. The two signals may be superimposed, as shown in Fig. 8. This figure also demonstrates how the observed signal originates. The amplitude of the H signal when the V signal is zero is given by point E_1. The amplitude of the V signal when the H signal is zero is given by point E_2. The ratio of E_1 to E_2 can be found by measuring the dB difference between the two points. Also, the ratio of the amplitudes E_3 and E_4 may be found in a similar fashion when the E makes 45° and 135° with the H antenna. The ratios E_1/E_2 and E_3/E_4 provide sufficient information to determine the polarization angle β and the ratio of semiminor axis to semimajor axis which determines $\cos\psi$. To facilitate this calculation, a modified projection of a Poincaré sphere may be employed.[5] Fig. 9(a) shows such a chart. The dB ratio of E_1/E_2 may be entered on the horizontal axis and E_3/E_4 on the vertical axis. In this manner, point M may be found. Using the overlay (Fig. 9(b)), the look angle ψ and polarization angle β may be read directly. It should be pointed out that the angle β is not just an effect of the orientation of the plane of tumble with respect to the plane of the antenna, but also includes Faraday rotation. By determining the space-fixed orientation of the plane of rotation from equation (1),

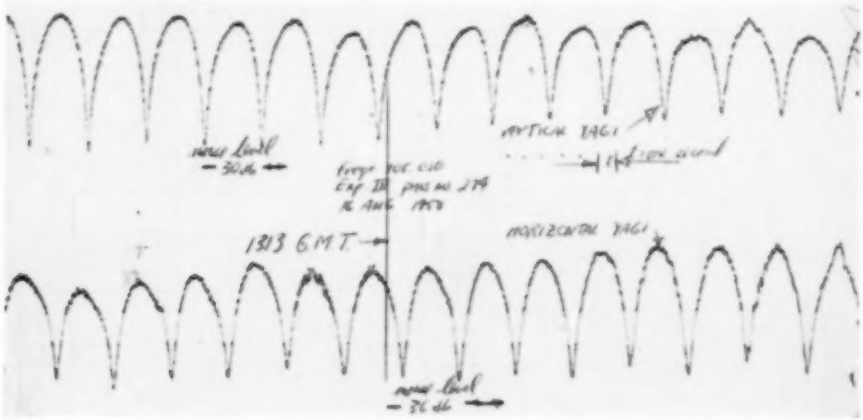


Fig. 7.

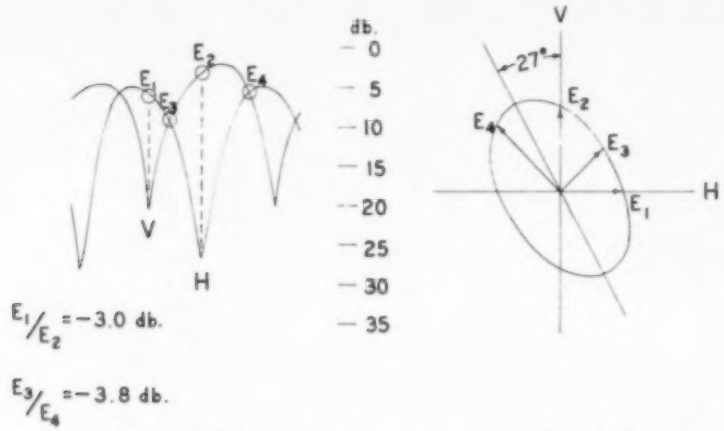


Fig. 8. Polarization ellipse resulting from superposition of *H* and *V* signals.

it is possible to calculate the angle that the semi-major axis makes with the plane of the antenna. The difference in this calculated angle and the measured polarization angle β is the amount of Faraday rotation that the incoming signal has undergone. Therefore, this scheme will provide orientation measurements and a method for measuring Faraday rotation.

Another important aspect is the determination of the direction in space assumed by the longitudinal axis **A**, at any time t. This can be accomplished by observing the time at which the $|E|$ is a maximum. Since $|E|$ is a maximum when the dipole is perpendicular to the line of sight **R**, and since the dipole is always perpendicular to the angular momentum vector $\boldsymbol{\Omega}$, then at time of E_{max}, the longitudinal axis must be mutually orthogonal to both **R** and $\boldsymbol{\Omega}$. The direction in space assumed by the dipole can be found by $\mathbf{R} \times \boldsymbol{\Omega}$. Knowing the position of the longitudinal axis at a particular time, together with knowledge of the rotation rate and the orientation of the plane in which **A** is constrained to move, it is possible to determine the orientation of **A** at any time desired.

RESULTS

Some Explorer IV data that permitted use of

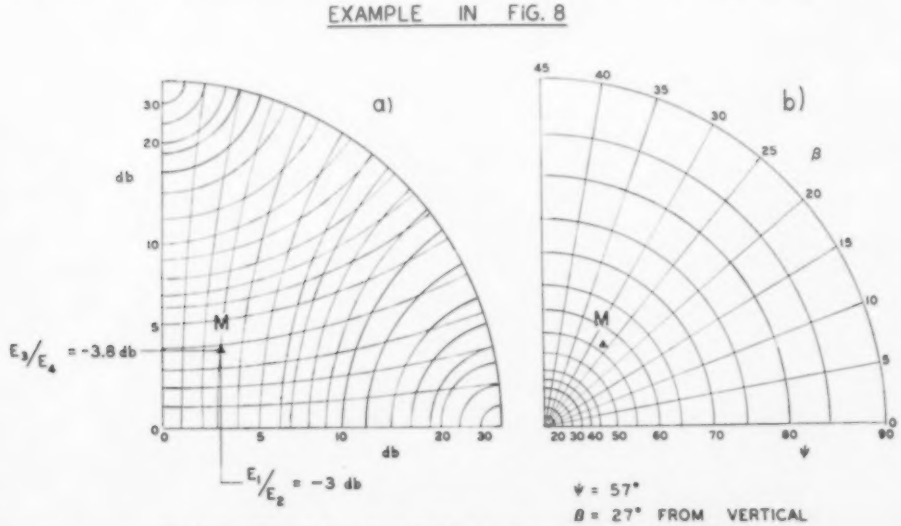


Fig. 9. Modified Poincaré projection used to solve example in Fig. 8.

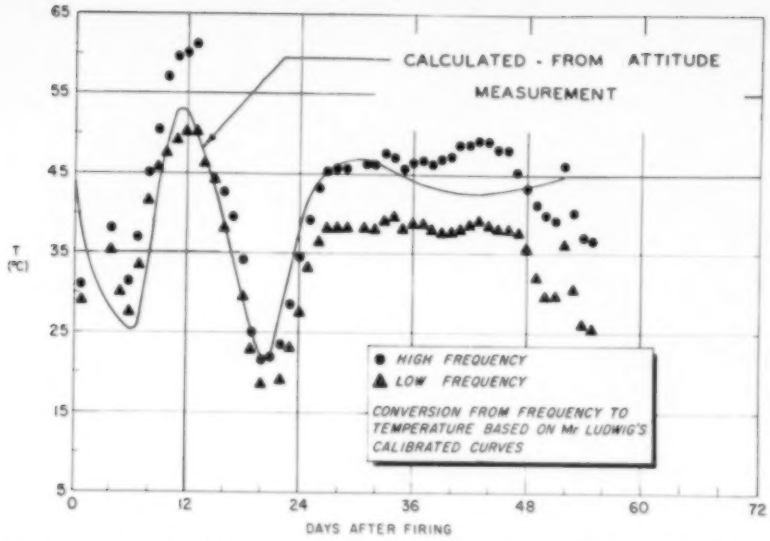


Fig. 10. Temperature of Explorer IV based on channel three frequencies received by ABMA tracking station.

the technique described previously was available from Ft. Monmouth. However, there is not sufficient data of this sort to give a complete history of the orientation of the satellite. It is, therefore, necessary to rely on the "ball park" orientation provided by the Army Ballistic Missile Agency and Stanford Research Institute data, with the Ft. Monmouth data providing a consistent check with the results obtained.

Using the orientation results shown in Fig. 6, it was possible to explain the temperature fluctuation described earlier. The solar aspect angle was found from the orientation vector and the Sun line. From this, the time average area presented to the Sun was determined and, thus, the temperature could be calculated. Fig. 10 shows the predicted temperature plotted with the measured temperature taken from the two sub-carrier oscillator frequencies. It may be seen that the two compare quite favorably.

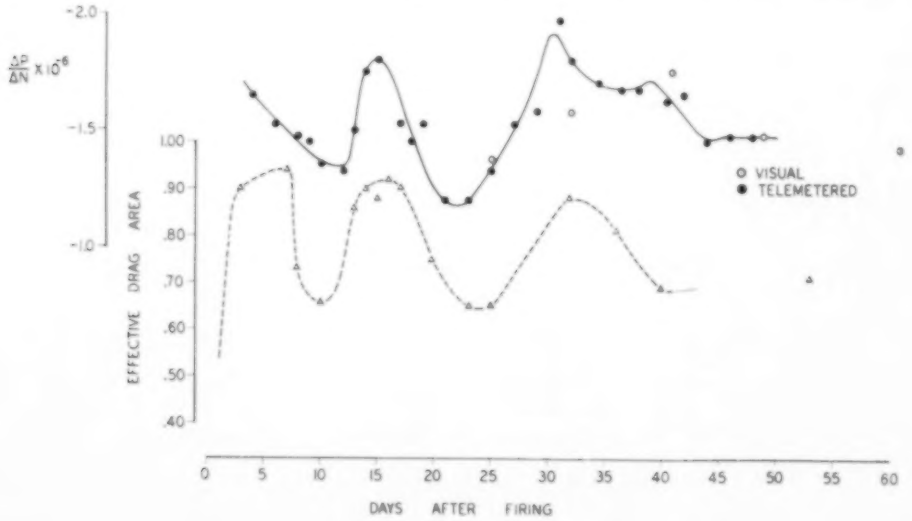


Fig. 11. Effective drag area and drag forces vs. days after firing.

Figure 11 shows the rate change of period resulting from drag forces during the transmitter lifetime, as determined by Smithsonian Astrophysical Observatory.[6] Along with this plot is the calculated area presented to the perigee velocity vector. It may be seen that the drag forces essentially vary as the presented area.

Additional work is being carried out in measuring Faraday rotation, correlating radiation counting with orientation measurements to determine the spatial distribution of the observed radiation, and determining the nature of the precessing torques.

CONCLUSIONS

Several important points may be concluded from this study:

(1) We have shown that orientation measurement is feasible using R.F. techniques and have outlined a method whereby the angular momentum vector and the direction of the longitudinal axis may be defined at any time. The discussion has assumed a tumbling dipole, but with certain modifications it may be extended to almost any type of satellite.

(2) Having defined the orientation of the satellite, we may make use of this knowledge to provide a better interpretation of measured data.

(3) We have demonstrated that body torques exist in the case of Explorer IV that can radically change the direction, but not the magnitude of the angular momentum. The nature of these torques is still not known. Since the angular momentum remains almost constant in magnitude for two months, the torques do not appear dissipative and, hence, are probably not magnetic. Aerodynamic torques arising from a couple of about 10 cm between the center of aerodynamic pressure and the center of gravity could account for precessions of the observed magnitude *if* they were always applied in a manner so as to be additive. However, due to the tumbling motion, any aerodynamic torques would appear to time average out with little or no net effect. The gradient of the gravitational field can give rise to torques of the order of 30 dyne-cm. A precession of 10°/day would require about 100 dyne-cm. This precession requires more detailed study before any definite conclusions can be reached.

(4) Due to the combined tumble and precession of long cylindrical satellites, a detector looking along the longitudinal axis will sweep out the entire celestial sphere in a short time. This, along with the slow tumble rate and the possibility for performing orientation measurements, makes such a configuration an ideal vehicle for carrying out ionospheric measurements, as well as gamma, ultraviolet, visual and radio astronomy or any other scientific experiment where it is desirable to measure the directionality of an effect.

Acknowledgments—I wish to express my appreciation to Mr. Gerhard Heller, Dr. Charles Lundquist, Dr. Russell Shelton and Mr. William Snoddy, of Research Projects Laboratory, ABMA, for their help and suggestions in this study, and also to Miss Joan Kassner and Mr. George Prince of the Computation Laboratory, ABMA, for the numerous calculations on the IBM 704 and Pace Analog Computers.

REFERENCES

1. G. Heller, *Thermal Control of the Explorer Satellites*, ARS Semi-Annual Meeting, San Diego, Calif. (1959).
2. W. C. Snoddy, *Theoretical and Measured Temperatures for Explorer IV*, ABMA Tech. Note DV-TN-6-59 (1959).
3. R. J. Naumann, *Determination of Satellite Spatial Orientation*, presented at the Army Science Conference, West Point, N.Y. (1959).
4. R. J. Naumann, *Attitude Determination of Explorer IV* ABMA Tech. Note DV-TN-19-59 (1959).
5. *Reference Data for Radio Engineers*, p. 669, International Telephone and Telegraph Corporation, New York (1956).
6. Personal communication with Dr. Whitney and Dr. Lautman, Smithsonian Astrophysical Observatory, Cambridge, Mass. (1959).

REVIEW OF THE ARDC MODEL ATMOSPHERE, 1959

K. S. W. CHAMPION, R. A. MINZNER* and H. L. POND
Air Force Cambridge Research Center, Laurence G. Hanscom Field

Abstract—The ARDC Model Atmosphere, 1959 constitutes a revision of the ARDC Model Atmosphere, 1956, made necessary by the accumulation of considerable new experimental upper atmosphere data. All available rocket and satellite density data up to 700 km were assembled and critically evaluated and the best average values of density were chosen. A molecular-scale temperature curve was devised which is consistent with the density curve. To obtain the corresponding kinetic temperatures, a curve of mean molecular weight as a function of altitude was specified.

I. INTRODUCTION

When the 1956 ARDC Model Atmosphere[1] was developed, the available experimental data extended to approximately 150 km. Thus it was not surprising that when satellite density data up to about 650 km became available, it was not in good agreement with the curve extrapolated from the earlier data. This indicated the necessity of making a major revision of the 1956 Model Atmosphere. The 1959 Model is the product of this revision. First, densities were calculated from orbit data of each satellite using a formula developed by Sterne[2] which gives the density at the perigee of the orbit. The method of calculation and the assumptions involved were investigated critically.[3] Second, all available rocket density data were collected and critically evaluated. Although there probably are seasonal and locational variations in density, most of the variation between measured values was found to be within the experimental accuracy.

A smooth curve was drawn through the average measured value of the density at each altitude. Up to 90 km this curve coincides with that of the 1956 Model; between 90 km and 153 km the new model lies slightly below the 1956 Model, but the curves cross and above 153 km the new model is higher. At an altitude of 350 km the new model is higher by a factor of 12 and at 660 km the factor is 24. The density curve and experimental data are plotted in Fig. 1.

By trial and error a molecular-scale temperature function was obtained which generates, via the hydrostatic equation, a density curve indistinguishable from that obtained empirically. This molecular-scale temperature function is taken to be the defining property of the 1959 Model Atmosphere.

II. THE MOLECULAR-SCALE TEMPERATURE

The concept of geopotential is introduced by the equation

$$H = \frac{1}{G}\int_0^Z g(Z)\mathrm{d}Z \qquad (1)$$

where

H is the geopotential altitude,
g is the acceleration of gravity,
G is a factor depending on the units used (9·80665) in the MKS system), and
Z is the geometric altitude.

The atmosphere is assumed to obey the perfect gas law and to be in hydrostatic equilibrium. The molecular-scale temperature, T_M, is defined by the equation

$$T/T_M = M/M_0, \qquad (2)$$

where

T is the absolute temperature,
T_M is the molecular-scale temperature,
M is the actual molecular weight, which is a function of altitude, and
M_0 is the sea-level value of the molecular weight.

* Now at the Geophysics Corporation of America, Bedford, Massachusetts.

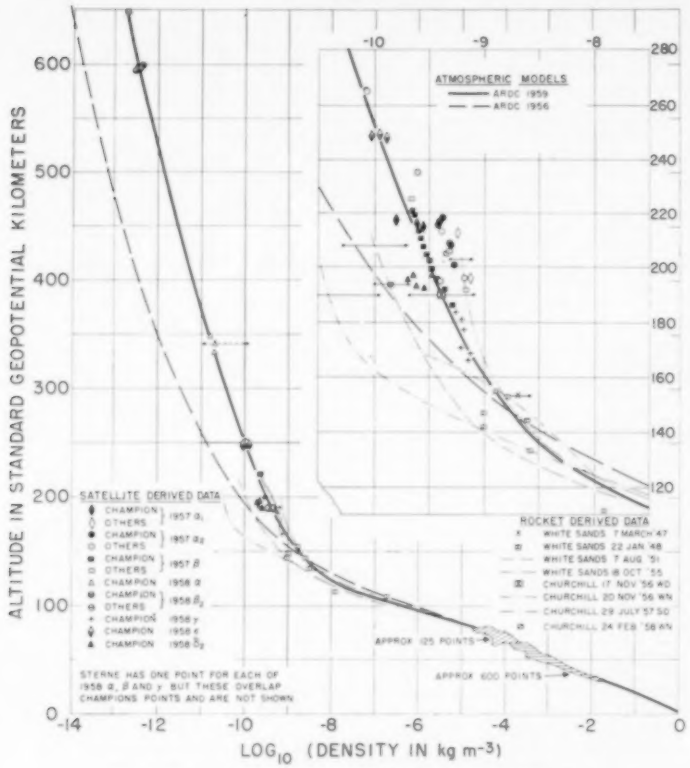

ALTITUDE IN STANDARD GEOPOTENTIAL KILOMETERS
LOG10 (DENSITY IN kg m-3)
ATMOSPHERIC MODELS
ARDC 1959
ARDC 1956
SATELLITE DERIVED DATA
ROCKET DERIVED DATA
APPROX 125 POINTS
APPROX 600 POINTS
STERNE HAS ONE POINT FOR EACH OF 1958 α, β AND γ BUT THESE OVERLAP CHAMPIONS POINTS AND ARE NOT SHOWN

Fig. 1. Variation of atmospheric density with altitude.

As in the 1956 ARDC Model, the molecular-scale temperature is assumed to be expressed in terms of H in the form

$$T_M = (T_M)_b + L_M(H - H_b) \tag{3}$$

where

$(T_M)_b$ is the value of T_M at altitude H_b, and
L_M is the slope of the molecular-scale temperature.

Thus, the molecular-scale temperature curve consists of a series of straight line segments. With T_M given in this form, the density is given by the equation

$$\rho = \begin{cases} \rho_b \left[\dfrac{(T_M)_b}{(T_M)_b + L_M(H - H_b)}\right]^{\left(1 + \frac{GM_e}{R^* L_M}\right)} & L_M \neq 0 \\ \rho_b \text{ exponential } \left(-\dfrac{GM_0(H - H_b)}{R^*(T_M)_b}\right), & L_M = 0 \end{cases} \tag{4}$$

where

R^* is the universal gas constant,
ρ is the density at altitude H, and
ρ_b is the value of ρ at altitude H_b.

From equation (2), it follows that

$$dT_M/dH = A + B\, dT/dZ \tag{5}$$

where

$$A = T\frac{d(M_0/M)}{dH}$$

and

$$B = \frac{M_0}{M}\frac{dZ}{dH}$$

The geopotential altitude can be approximated by

$$H = \frac{g_0}{G}\frac{Z}{1 + (Z/R)} \tag{6}$$

where

g_0 is the sea-level value of the acceleration of gravity, and
R is the effective radius of the Earth (taken as 6356·766 km). Ref. 1, appendix M.

Then the expression for B can be written as

$$B = \frac{G}{g_0}\frac{M_0}{M}\left(1+\frac{Z}{R}\right)^2. \qquad (7)$$

While the mean molecular weight at great altitudes is not known, some reasonable assumptions can be made in order to estimate A and B.

Table 1. Molecular Weight Values

Z_{km}	H_{km}'	M	$\frac{M_0}{M}$	$\frac{\Delta(M_0/M)}{\Delta H}$
90	88·7	29	1·0	
				0·0016
600	548·0	17·5	1·7	
				0·0004
1000	864·0	16·2	1·8	

As an example, consider the values given in Table 1.

If it is now assumed that at an altitude of 600 km the kinetic temperature, T, is between 1500° and 2000°K and $d(M_0/M)/dH$ is taken as 0·001, then

$$1{\cdot}5 < A < 2{\cdot}0$$
$$B \doteq 2$$

Thus, the relation between the slope of the molecular-scale temperature curve and the slope of the kinetic temperature curve is given by equation (5), and at 600 km both A and B are approximately equal to two.

Figure 2a shows three molecular-scale temperature curves, each with different slopes above 150 km′ but yielding the same density at 600 km′. The first of these is isothermal at 2075°K for altitudes above 280 km′. The second, Model II, has a positive slope of 3·5°K/km′ for altitudes above 200 km′. The third, Model III, has a positive slope of 5·5°K/km′ for altitudes above 150 km′. The corresponding density curves are shown in Fig. 2b. It is apparent that all three models yield nearly the same densities below 700 km′. The greatest difference between these densities is about a factor of 3 which is comparable to the errors in the satellite density

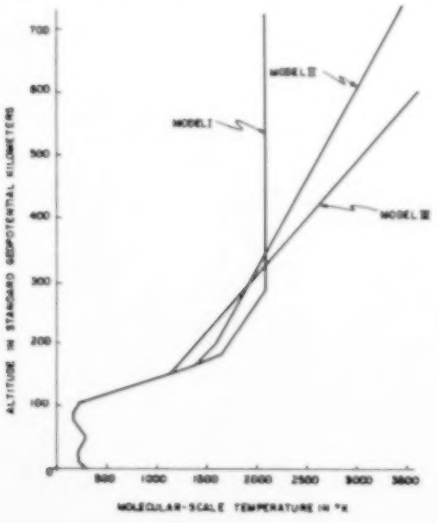


Fig. 2a.

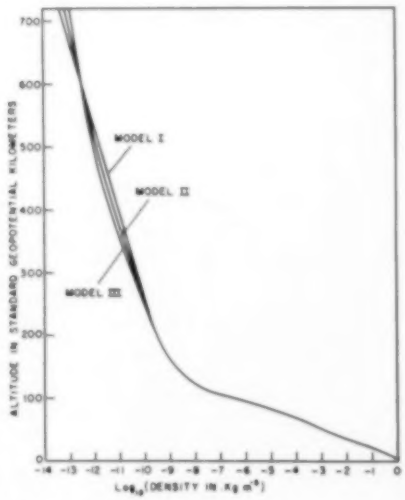


Fig. 2b.

data. Hence, any one of these functions could be considered as a suitable fit to the data between 150 and 700 km′. If these three models are extended to 1000 km′, however, Models I and III differ in density by a factor of nearly 30. The middle curve, Model II, gives the density for the 1959 Model Atmosphere. With the assumed mean molecular weight curve (shown in Fig. 3), the slope of the kinetic temperature curve at an altitude of 600 km is found to be 1·25°K/km.

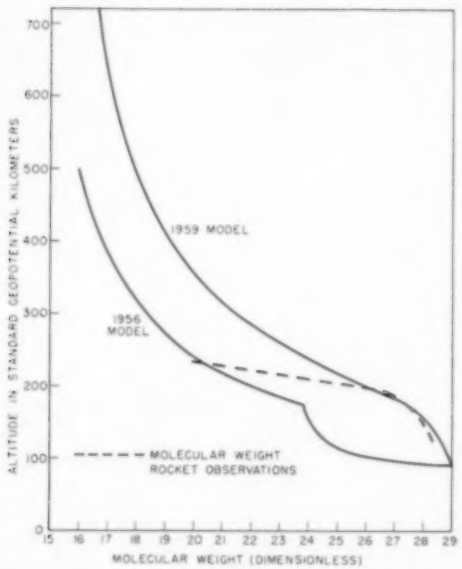


Fig. 3. Variation of mean molecular weight with altitude.

Chapman has used a mean slope of 4°K/km for the kinetic temperature curve at altitudes above 500 km.[4, 5] Such a slope yields with linear extrapolation (taking $T = 2000$°K at $Z = 600$ km) a temperature of the order of 10^5°K at an altitude of about 4 Earth radii, which is also in agreement with estimates made by Chapman.[4] A slope of 4°K/km for the kinetic temperature curve would require, according to the above analysis, a slope of about 10°K/km′ at 600 km for the molecular-scale temperature curve. Thus, it is possible that the slope should be greater than that used in the 1959 ARDC Model Atmosphere. However, it should be remembered that the slope of 3·5°K/km′ used for the molecular-scale temperature in the 1959 Model lead to a density curve that is in excellent agreement with the available data for altitudes between 200 and 600 km.

III. MOLECULAR WEIGHT

To proceed from the molecular-scale temperature curve to a curve of kinetic temperature, it is necessary to know the mean molecular weight as a function of altitude. A mean molecular-weight curve may be based on a theoretical description of the atmosphere or on the results of experimental probes. Unfortunately, at the present time neither basis is adequate to an altitude of 700 km, although it is hoped that satellite mass spectrometric data will soon remedy this deficiency.

For the 1956 Model Atmosphere it was assumed that the sea-level value of mean molecular weight of 28·97 did not change significantly up to 90 km′. The variation of mean molecular weight with altitude above 90 km′ used in the model was based on the theoretical calculations of Miller[6]. These calculations indicated that oxygen dissociation commences sharply at 90 km′ and is nearly complete by 175 km′, where diffusive equilibrium applies.

Recent rocket data reported by Townsend[7] suggest that the molecular weights of the 1956 ARDC Model are in error, particularly between 90 and 200 km′. The average molecular weights deduced from rocket flights of a Bennett mass spectrometer are compared with the 1956 Model values in Fig. 3. However, molecular weights determined from this curve, when combined with the values of molecular-scale temperature at corresponding altitudes, yielded a kinetic temperature curve with negative gradients ($dT/dH < 0$) in the region just above 200 km′. Townsend's values of M are based on measurements which are probably much less reliable than the various density data. Furthermore, the molecular-weight data above 180 km′ represent the results of only a single rocket flight, while at lower altitudes there were three flights. Accepting the molecular-weight data below 200 km′ and keeping the restriction of $dT/dH \geqslant 0$ then leads to the molecular weight curve for the 1959

Model shown in the figure. The equation of this curve is:

$$M = \begin{cases} 28{\cdot}966 & , -5\,\text{km}' \leqslant H \leqslant 90\,\text{km}' \\ 22{\cdot}000 - 5{\cdot}0448 \arctan \left[\dfrac{H-220}{25}\right] & , 90\,\text{km}' \leqslant H \leqslant 180\,\text{km}' \\ 27{\cdot}106 - 7{\cdot}9357 \arctan \left[\dfrac{H-180}{140}\right] & , 180\,\text{km}' \leqslant H \end{cases} \quad (8)$$

The constants multiplying the arctangent functions were determined after the other constants had been selected and were used to adjust the different segments at junction and end points.

The mean molecular weight curve for the 1959 Model results from an attempt to make use of the apparently best available experimental data. It leads to such questions as: Does oxygen dissociation commence at a higher altitude than 90 km, or does it start at 90 km but increase with altitude at a much slower rate than was previously believed? The apparent alternative is to reject the mass spectrometer measurements between 100 and 200 km. It is hoped that further experiments and theoretical work will yield conclusive answers to these questions.

The kinetic temperature of the atmosphere has been measured directly up to balloon altitudes but it is difficult to measure ambient temperature with rocket probes, due to temperature perturbations introduced by the high velocity missile. At altitudes less than 100 km, grenades have been used to measure the velocity of sound and, hence, temperature but this is an integrated value. At higher altitudes kinetic temperature has been obtained solely by deduction from measurements of other physical properties. Thus there is an urgent need for direct temperature measurements above 100 km using suitably designed probing techniques.

IV. THE TRANSPORT PROPERTIES

The properties that will be considered in this section are the coefficients of viscosity, kinematic viscosity, thermal conductivity, and the velocity of sound. These properties are functions of the kinetic or molecular-scale temperature, density and the ratio of specific heats of the gas. The values of temperatures and densities are known up to 700 km with reasonable accuracy and are shown in British units in Fig. 4. The ratio of specific heats (γ), which is involved in the velocity of sound, is a function of the composition, including atomic or molecular state, of the atmosphere. To tabulate values of the velocity of sound at altitudes above 90 km, it would first be desirable to tabulate values of γ at these altitudes. This has not been done in the present model.

The next question to be raised is the accuracy at higher altitudes of the formulae for the transport properties. The formulae used in the 1959 Model for viscosity and thermal conductivity are both empirical, and assume the sea-level composition of the air. Sutherland's formula for viscosity is calculated assuming that the molecules are rigid attracting spheres and can be written in the form

$$\mu = \beta T^{3/2}/(T + S) \quad (9)$$

where β and S are constants whose values depend on the gas. With suitable empirical values of β and S, equation (9) represents closely the variation of μ with temperature for a number of gases, including nitrogen and air of sea-level composition. The agreement with experimental data is better at higher temperatures (e.g. 250°C) which is a desirable feature. However, the effects of changing composition, dissociation and ionization, as encountered at high altitudes, on equation (9) are largely unknown.

Theoretically, the coefficient of viscosity is independent of density, and this is true as long as the mean free path is much less than the dimensions considered. From the point of view of the circulation of the air, this will apply up to about 600 km. Above this altitude, where the mean free path is about 90 km, the properties of the atmosphere vary significantly over a vertical distance of one mean free path. From the point of view of interaction with a missile or other object, the maximum altitude would be somewhat lower, e.g. 100 km where the free path is 0·2 m. When the free path is not small compared with the dimensions (e.g. of a missile),

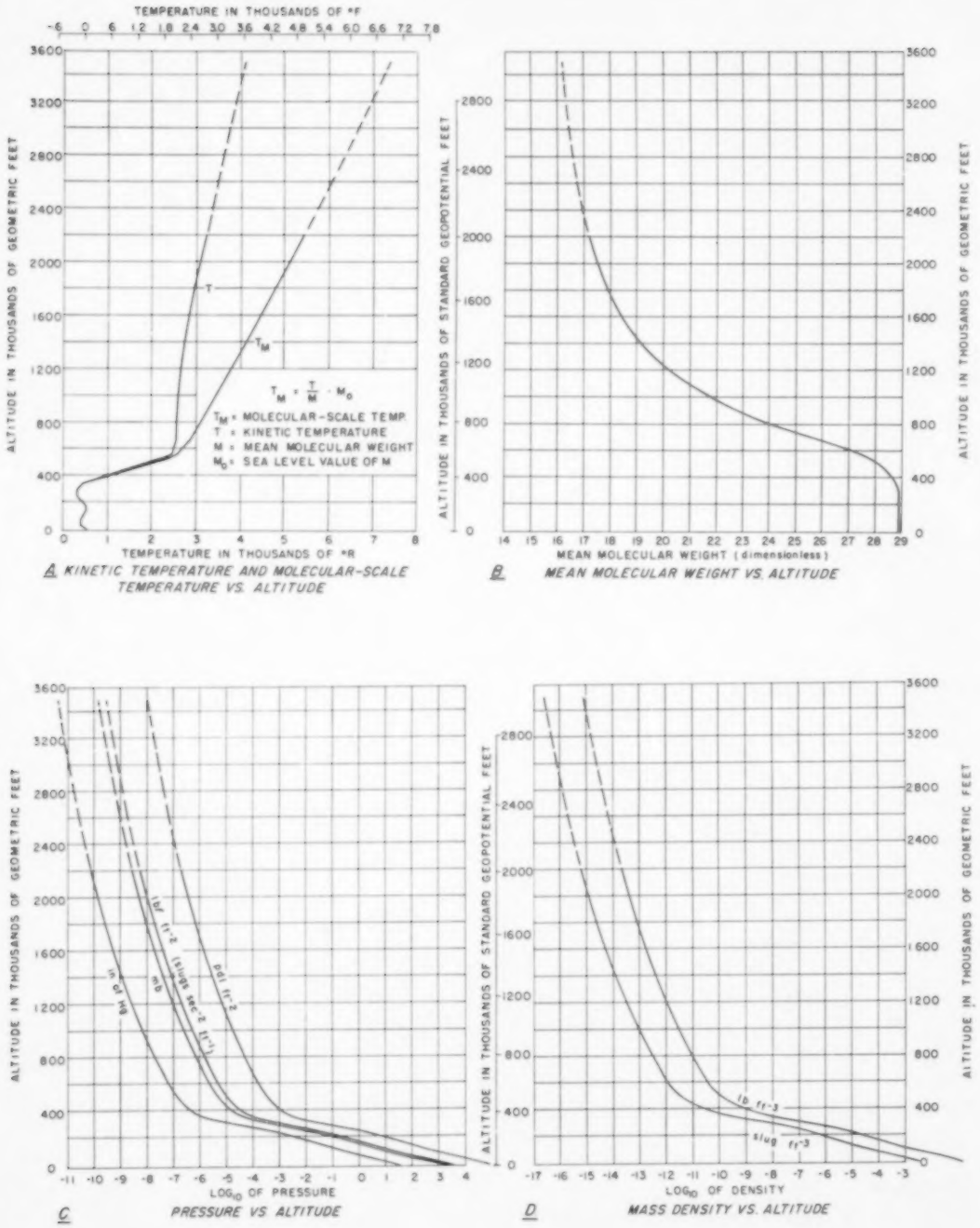

TEMPERATURE IN THOUSANDS OF °F
ALTITUDE IN THOUSANDS OF GEOMETRIC FEET
TEMPERATURE IN THOUSANDS OF °R
T
T_M
T_M = (T/M) · M_0
T_M = MOLECULAR-SCALE TEMP.
T = KINETIC TEMPERATURE
M = MEAN MOLECULAR WEIGHT
M_0 = SEA LEVEL VALUE OF M
A KINETIC TEMPERATURE AND MOLECULAR-SCALE TEMPERATURE VS. ALTITUDE
ALTITUDE IN THOUSANDS OF STANDARD GEOPOTENTIAL FEET
MEAN MOLECULAR WEIGHT (dimensionless)
B MEAN MOLECULAR WEIGHT VS. ALTITUDE
LOG10 OF PRESSURE
C PRESSURE VS. ALTITUDE
LOG10 OF DENSITY
lb ft-3
slug ft-3
D MASS DENSITY VS. ALTITUDE

Fig. 4. A: Kinetic temperature and molecular-scale temperature vs. altitude; B: mean molecular weight vs. altitude; C: pressure vs. altitude; D: mass density vs. altitude.

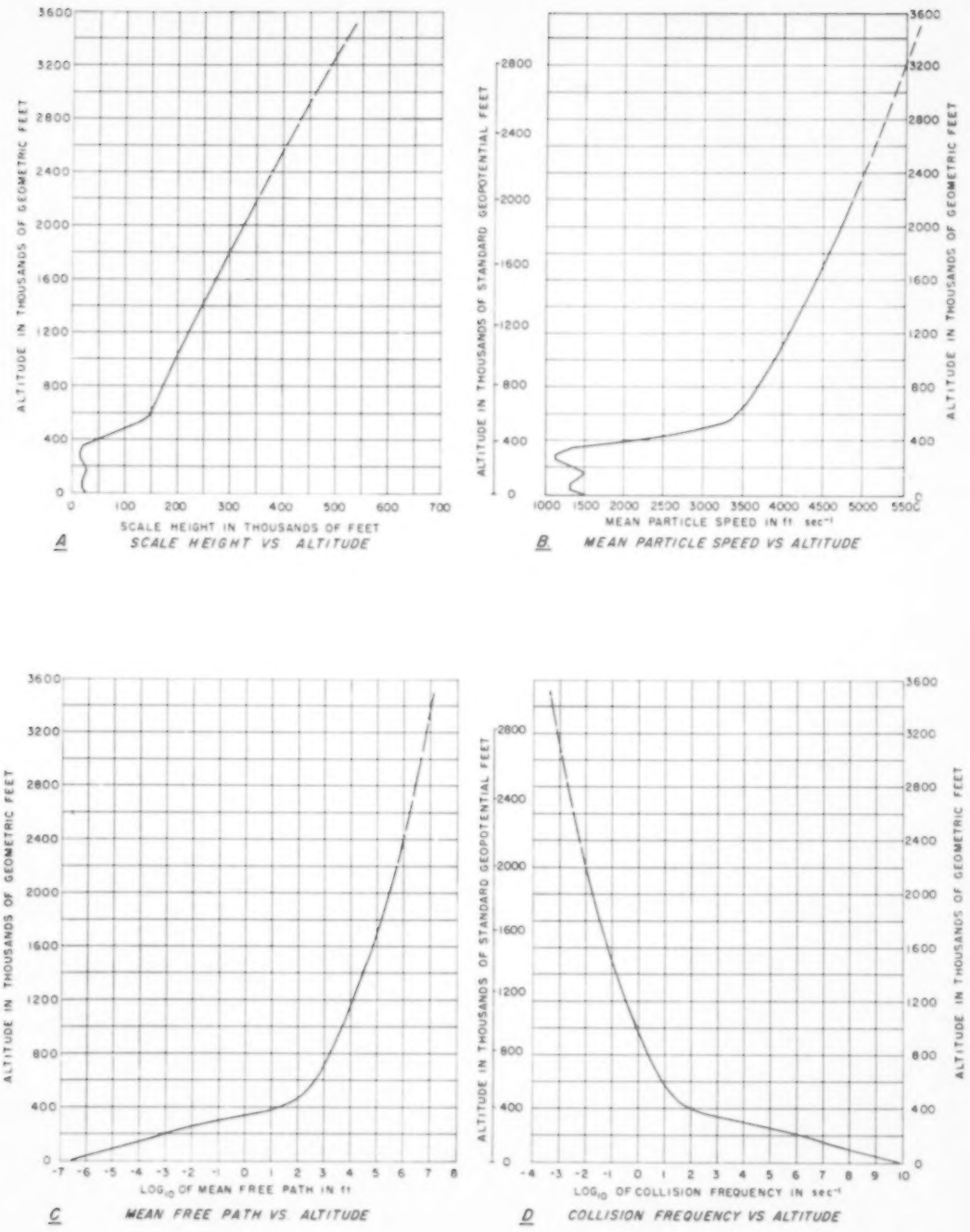

ALTITUDE IN THOUSANDS OF GEOMETRIC FEET
SCALE HEIGHT IN THOUSANDS OF FEET
A SCALE HEIGHT VS ALTITUDE
ALTITUDE IN THOUSANDS OF STANDARD GEOPOTENTIAL FEET
MEAN PARTICLE SPEED IN ft sec⁻¹
B MEAN PARTICLE SPEED VS ALTITUDE
LOG10 OF MEAN FREE PATH IN ft
C MEAN FREE PATH VS. ALTITUDE
LOG10 OF COLLISION FREQUENCY IN sec⁻¹
D COLLISION FREQUENCY VS ALTITUDE

Fig. 5. A: Scale height vs. altitude; B: mean particle speed vs. altitude; C: mean free path vs. altitude; D: collision frequency vs. altitude.

there will be slipping of gas at the surface and the apparent value of the viscosity will be smaller. In addition, the apparent viscosity observed by a missile travelling at high velocity when the air flow is turbulent (for example near or above Mach 1) is a function of its velocity.

The formula used for thermal conductivity

$$K = \frac{6{\cdot}325 \times 10^{-7} T^{3{\cdot}2}}{T + 245{\cdot}4 \times 10^{-12/T}} \qquad (10)$$

is an empirical relation quoted by Hilsenrath *et al.*[8]

Using Sutherland's model it can be shown hat

$$K = \frac{ac_v T^{3{\cdot}2}}{T + 1{\cdot}03\,S + (0{\cdot}03\,S^2/T)} \qquad (11)$$

where a is a constant and c_v is the specific heat at constant volume of the air. This is not in exactly the same form as equation (10), but both are approximations to the temperature dependence of actual gases.

At low pressures, for which the mean free path becomes comparable to the dimensions of an object or missile, or such that the temperature gradient changes appreciably in a mean free path, significant deviations from the above expressions occur. Again, the effects of changing composition, dissociation and ionization, as encountered at high altitudes, on these expressions for thermal conductivity are largely unknown. Spitzer[9] considers the thermal conductivity of a fully ionized gas and it should be

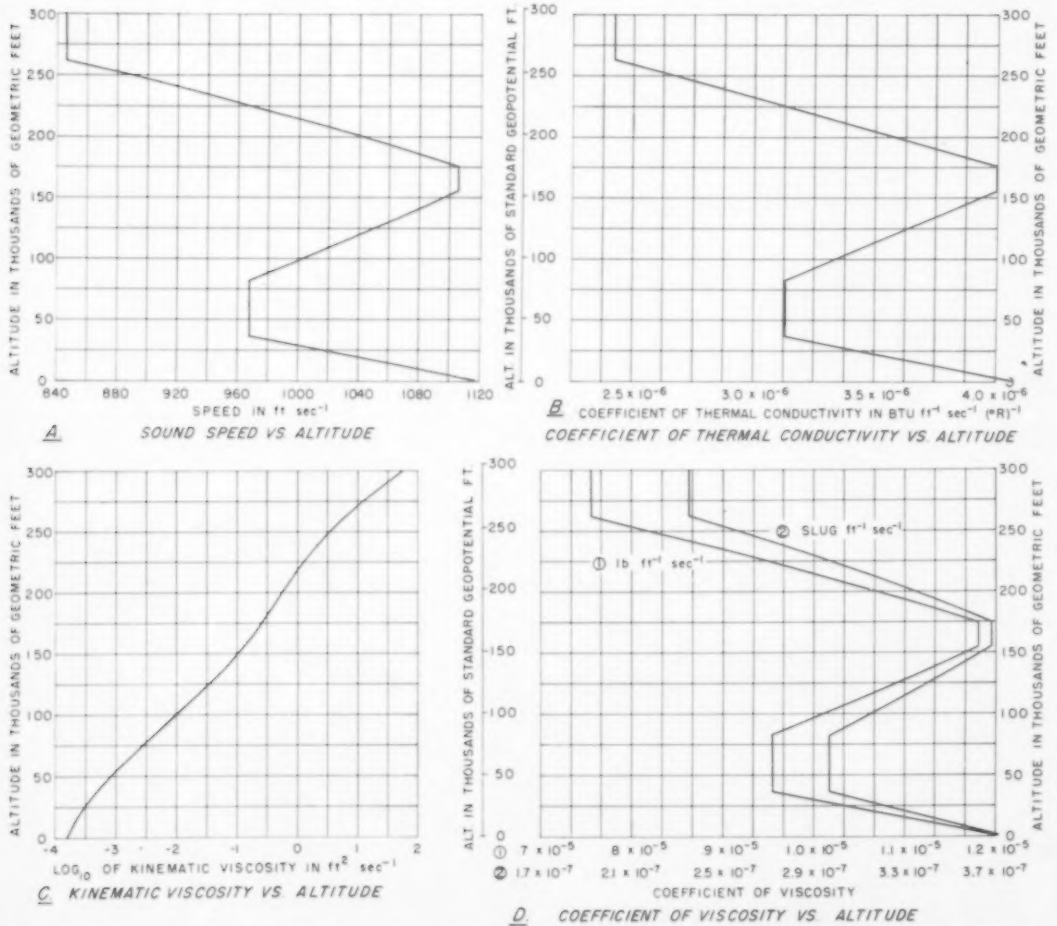

ALTITUDE IN THOUSANDS OF GEOMETRIC FEET
SPEED IN ft sec⁻¹
A SOUND SPEED VS. ALTITUDE
ALT IN THOUSANDS OF STANDARD GEOPOTENTIAL FT
COEFFICIENT OF THERMAL CONDUCTIVITY IN BTU ft⁻¹ sec⁻¹ (°R)⁻¹
B COEFFICIENT OF THERMAL CONDUCTIVITY VS. ALTITUDE
LOG10 OF KINEMATIC VISCOSITY IN ft² sec⁻¹
C KINEMATIC VISCOSITY VS. ALTITUDE
① lb ft⁻¹ sec⁻¹
② SLUG ft⁻¹ sec⁻¹
COEFFICIENT OF VISCOSITY
D COEFFICIENT OF VISCOSITY VS. ALTITUDE

Fig. 6. A: Sound speed vs. altitude; B: coefficient of thermal conductivity vs. altitude; C: kinematic viscosity vs. altitude; D: coefficient of viscosity vs. altitude.

possible to derive the conductivity of a partially ionized gas. It is interesting to note that in a gas containing ions, a temperature gradient sets up a thermoelectric field which acts to reduce the flow of heat.

The concept of the speed of sound is related to the attenuation of sound transmission in that as the intensity approaches zero, the concept of speed of transmission becomes meaningless. The rate of absorption or attenuation of sound energy per unit length in air is related to frequency of the sound and the air pressure so that the attenuation increases with increasing frequency and also increases with decreasing pressure. Thus, while the sound-transmission efficiency over a given distance approaches zero for very high frequencies at sea-level pressures, it also approaches zero for very low frequencies at the low pressures of the upper atmosphere, thereby suggesting an upper limit for tabulating sound velocity.

Plots of the four quantities discussed in this section are given in Fig. 6 up to 300,000 ft (approximately 90 km). In Fig. 5 are given mean particle speed, mean free path and collision frequency up to 3,000,000 ft. It should be noted that, at extreme altitudes, the effects of collisions between ions, which form an appreciable part of the gas, will result in some reduction in the plotted mean free path and increase in collision frequency, due to their large Coulomb cross sections.

CONCLUSION

This paper has presented, in the form of figures, the major new data to be included in the 1959 ARDC Model Atmosphere. Most of the data has been plotted in British units so that it should be of immediate value to design engineers who require values of the properties of the upper atmosphere.

The theoretical and experimental bases of the calculations have been outlined. The limitations of some of the theoretical formulae have been critically discussed. In addition, the regions where experimental data are most urgently needed have been pointed out. Finally, it can be said that it is believed that the data in the 1959 Model is substantially more accurate, at higher altitudes, than that in the 1956 Model but the authors are looking forward to more accurate theory and experiments to yield a still more accurate model in the future.

REFERENCES

1. R. A. Minzner and W. S. Ripley, *The ARDC Model Atmosphere 1956*, AFCRC TN-56-204, ASTIA Document 110233 (1956).
2. T. E. Sterne, *Science* **127**, 1245 (1958).
3. K. S. W. Champion, *Atmospheric Densities from Satellite Orbit Data and the 1959 ARDC Model Atmosphere*, to be published.
4. S. Chapman, *Smithson. Contr. Astrophys.* **2**, 1 (1957).
5. S. Chapman, *The Threshold of Space* (ed. by M. Zelikoff), Pergamon Press, New York (1957).
6. L. E. Miller, *J. Geophys. Res.* **62**, 351 (1957).
7. J. W. Townsend, *Composition of the Upper Atmosphere*, unpublished manuscript, U.S. Naval Res. Lab. (Oct. 1958).
8. J. Hilsenrath *et al.*, *Nat. Bur. Stand. Circ.* No. 564 (1955).
9. L. Spitzer, *Physics of Fully Ionized Gases, Interscience Tracts on Physics and Astronomy*, Vol. 3, Interscience Publishers, Inc., New York (1956).

END OF VOL.